PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-272000

(43)Date of publication of application : 13.10.1998

(51)Int.Cl. C12Q 1/68
C12N 15/09
G06F 17/30
// G01N 33/50

(21)Application number : 09-369833

(22)Date of filing : 11.12.1997

(71)Applicant : AFFYMETRIX INC

(72)Inventor : WEBSTER TERESA A
MORRIS MACDONALD S
MITTMANN MICHAEL P
LOCKHART DAVID J
HO MING-HSIU
BERNHART DEREK
JEVONS LUIS C

(30)Priority

Priority number : 96 33053 Priority date : 12.12.1996 Priority country : US

97 828952 28.03.1997

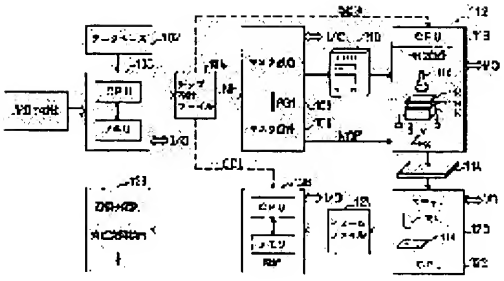
US

(54) ANALYTIC TECHNIQUE OF BIOLOGICAL SEQUENCE UTILIZING COMPUTER

(57)Abstract:

PROBLEM TO BE SOLVED: To analyze a biological sequence of nucleic acid, etc., with computer system by inputting and analyzing plural base calls to each base position along a part of nucleic acid sequence of a specimen and producing and displaying a single base call.

SOLUTION: In a method for analyzing a nucleic acid sequence of a specimen in a computer system, plural base calls to each base position along at least a part are inputted to a computer 100 and a base call having highest occurrence frequency among these plural base calls is obtained by database 102 and a single base call is produced as a base call having highest frequency of occurrence and inputted to a chip design file 104 and a pattern is arranged on a mask 110 and the chip design file 104 and an image file 124 are analyzed and single base call each led from plural base calls to a specific base position in the nucleic acid sequence of the specimen is displayed as an output 128.



LEGAL STATUS

[Date of request for examination] 22.05.2002

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application

converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号

特開平10－272000

(43)公開日　平成10年(1998)10月13日

(51)Int.Cl.⁶		識別記号	F I		
C12Q	1/68		C12Q	1/68	A
C12N	15/09		G01N	33/50	P
G06F	17/30		C12N	15/00	A
// G01N	33/50		G06F	15/40	370F

審査請求　未請求　請求項の数35　ＦＤ　外国語出願　（全 104 頁）

(21)出願番号　特願平9－369833

(22)出願日　平成９年(1997)12月11日

(31)優先権主張番号　６０／０３３，０５３

(32)優先日　1996年12月12日

(33)優先権主張国　米国（ＵＳ）

(31)優先権主張番号　０８／８２８，９５２

(32)優先日　1997年３月28日

(33)優先権主張国　米国（ＵＳ）

(71)出願人　598012740
アフィメトリックス・インコーポレーテッド
ＡＦＦＹＭＥＴＲＩＸ　ＩＮＣＯＲＰＯＲＡＴＥＤ
アメリカ合衆国　カリフォルニア州95051　サンタ・クララ，セントラル・エクスプレスウェイ，3380

(72)発明者　テレサ・エー．　・ウェブスター
アメリカ合衆国　カリフォルニア州94021　ローマ・マー，ペスカドーレ・ロード，10002

(74)代理人　弁理士　五十嵐　孝雄　（外２名）

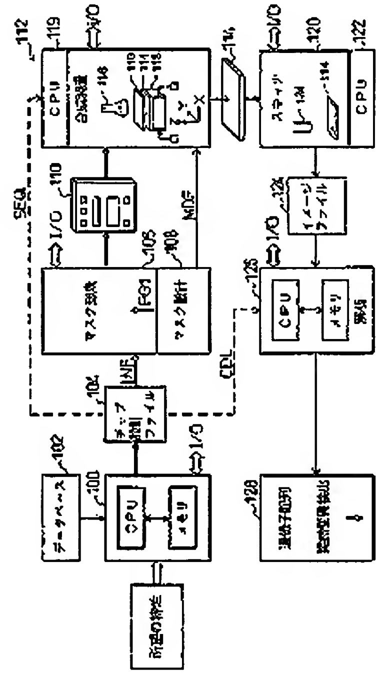
最終頁に続く

(54)【発明の名称】　コンピュータを利用した生物学的配列の解析技術

(57)【要約】

【課題】　核酸のような生物学的配列を解析するためのコンピュータ支援技術を提供する。

【解決手段】　コンピュータシステムは、核酸プローブと検体核酸配列との間のハイブリッド形成親和性を示すハイブリッド形成強度を解析して、検体配列における塩基を呼び出す（塩基コールを行う）。複数の塩基コールを組み合わせることにより、単一の塩基コールを形成する。また、コンピュータシステムは、ハイブリッド形成強度を解析することによって、遺伝子発現を監視したり、あるいは、基準からの遺伝子発現の変化を監視したりする。



1

【特許請求の範囲】

【請求項１】　コンピュータシステムにおいて．検体の核酸配列を解析するための方法であって，

前記検体の核酸配列の少なくとも一部分に沿った各塩基位置に対して．複数の塩基コールを入力する工程と，

各塩基位置に対して、前記複数の塩基コールを解析して単一塩基コールを生成する工程と，

前記検体の核酸配列の少なくとも一部分に沿った複数の塩基位置に対する複数の単一塩基コールであって．前記検体の核酸配列における特定の塩基位置に対する前記複数の塩基コールからそれぞれ導かれた単一塩基コールを表示する工程と．を備える方法。

【請求項２】　請求項１記載の方法であって，

前記解析工程は．

各塩基位置に対して、前記複数の塩基コールの中から最も発生頻度の高い塩基コールを求める工程と，

前記単一塩基コールを、前記塩基位置において最も発生頻度の高い塩基コールとして生成する工程と、を備える方法。

【請求項３】　請求項１記載の方法であって、更に、

ユーザーによる起動に応じて各塩基位置に前記複数の塩基コールを表示させるためのスクリーンアイコンを表示する工程、を備える方法。

【請求項４】　請求項１記載の方法であって、更に、

ユーザーによる起動に応じて各塩基位置に前記複数の塩基コールを表示させないためのスクリーンアイコンを表示する工程、を備える方法。

【請求項５】　請求項１記載の方法であって、更に、

塩基位置に従って、前記単一塩基コールと並べて各塩基における前記複数の塩基コールを表示する工程．を備える方法。

【請求項６】　請求項５記載の方法であって、更に、

前記複数の塩基コールの各々と共に．プローブと前記検体の核酸配列とのハイブリッド形成親和性を示すハイブリッド形成強度を表示する工程を備え．

各塩基コールは前記ハイブリッド形成強度の解析により求められる、方法。

【請求項７】　コンピュータシステムにおいて．検体の核酸配列における未知の塩基を呼び出すための方法であって．

複数組の核酸プローブに関するハイブリッド形成強度であって、それぞれ核酸プローブと前記検体の核酸配列との間のハイブリッド形成親和性を示すハイブリッド形成強度を入力する工程と，

各組のプローブに関して前記未知の塩基に対する塩基コールを演算する工程と，

前記複数組のプローブに関して最も発生頻度の高い前記未知の塩基に対する塩基コールに従って、前記複数組のプローブに対する単一塩基コールを演算する工程と、を備える方法。

2

【請求項８】　請求項７記載の方法であって、

各組のプローブが同一の参照配列に従って生成されたものである、方法。

【請求項９】　請求項７記載の方法であって、更に、

所定の条件下で．前記複数組の核酸プローブに関する前記単一塩基コールを特定する例外規則をチェックする工程．を備える方法。

【請求項１０】　コンピュータシステムにおいて．コンピュータによる塩基呼び出し処理に関するパラメータを動的に変更する方法であって、

ユーザによって変更可能なパラメータを含む前記塩基呼び出し処理を利用して、検体の核酸配列の少なくとも一部に関する塩基コールを生成する工程と，

前記検体の核酸配列の少なくとも一部に関する複数の塩基コールを表示する工程と．

前記塩基呼び出し処理のパラメータを表示する工程と，

前記塩基呼び出し処理のパラメータに対する新しい値を規定する入力をユーザーから受け取る工程と，

前記塩基呼び出し処理と前記パラメータに関する新しい値とを用いて．前記検体の核酸配列の少なくとも一部に関して更新された塩基コールを生成する工程と．

前記検体の核酸配列の少なくとも一部に関して前記更新された塩基コールを表示する工程と．を備える方法。

【請求項１１】　請求項１０記載の方法であって．更に．

前記塩基呼び出し処理のための、複数のユーザ変更可能なパラメータを表示する工程、を備える方法。

【請求項１２】　請求項１０記載の方法であって．

前記パラメータが、定数．閾値、および、範囲から成るグループから選択される．方法。

【請求項１３】　コンピュータシステムにおいて．検体の核酸配列における遺伝子の発現を監視する方法であって．

前記遺伝子に対して完全に相補的である完全対合プローブと前記遺伝子に対して不対合な塩基を少なくとも一つ備える不対合プローブとの複数のプローブ対に関する複数のハイブリッド形成強度であって．前記完全対合および不対合プローブと前記検体の核酸配列との間のハイブリッド形成親和性を示すハイブリッド形成強度を入力する工程と，

前記検体の核酸配列の遺伝子発現コールを生成するために．各対の完全対合プローブと不対合プローブのハイブリッド形成強度を比較する工程と，

前記遺伝子発現コールを表示する工程と、を備える方法。

【請求項１４】　請求項１３記載の方法であって．更に．

或る塩基位置における完全対合プローブと不対合プローブのハイブリッド形成強度の差分を．差分の閾値と比較する工程、を備える方法。

【請求項１５】　請求項１３記載の方法であって．更に．
或る塩基位置における完全対合プローブと不対合プローブのハイブリッド形成強度の比を、比の閾値と比較する工程．を備える方法。
【請求項１６】　請求項１３記載の方法であって．更に．
決定行列を用いて前記遺伝子発現コールを決定する工程．を備える方法。
【請求項１７】　請求項１３記載の方法であって．
前記遺伝子発現コールが．発現している、どちらともいえない、発現していない．から成るグループから選択される．方法。
【請求項１８】　コンピュータシステムにおいて．検体の核酸配列における遺伝子の発現を監視する方法であって．
前記遺伝子に対して完全に相補的である完全対合プローブと前記遺伝子に対して不対合な塩基を少なくとも一つ備える不対合プローブとの複数のプローブ対に関する複数のハイブリッド形成強度であって．前記完全対合および不対合プローブと前記検体の核酸配列との間のハイブリッド形成親和性を示すハイブリッド形成強度を入力する工程と、
各対の完全対合プローブと不対合プローブのハイブリッド形成強度を比較する工程と、
前記検体の核酸配列の遺伝子発現コールを生成する工程と．を備える方法。
【請求項１９】　請求項１８記載の方法であって．更に．
或る塩基位置における完全対合プローブと不対合プローブのハイブリッド形成強度の差分を．差分の閾値と比較する工程、を備える方法。
【請求項２０】　請求項１８記載の方法であって．更に．
或る塩基位置における完全対合プローブと不対合プローブのハイブリッド形成強度の比を、比の閾値と比較する工程．を備える方法。
【請求項２１】　請求項１８記載の方法であって．更に．
決定行列を用いて前記遺伝子発現コールを求める工程、を備える方法。
【請求項２２】　請求項１８記載の方法であって．
前記遺伝子発現コールが．発現している、どちらともいえない、発現していない．から成るグループから選択される．方法。
【請求項２３】　コンピュータシステムにおいて．検体の核酸配列における遺伝子の発現変化を監視する方法であって、
前記遺伝子に対して完全に相補的である完全対合プローブと前記遺伝子に対して不対合な塩基を少なくとも一つ

備える不対合プローブとの複数のプローブ対に関する複数のハイブリッド形成強度であって．前記完全対合および不対合プローブと前記検体の核酸配列との間のハイブリッド形成親和性を示すハイブリッド形成強度を入力する工程と、
前記検体の核酸配列の遺伝子発現レベルを生成するために．各対の完全対合プローブと不対合プローブのハイブリッド形成強度を比較する工程と、
前記遺伝子発現レベルを．基準の遺伝子発現レベルと比較することにより発現変化を求める工程と、
前記検体の核酸配列における前記遺伝子の発現変化を表示する工程と．を備える方法。
【請求項２４】　請求項２３記載の方法であって．
前記発現変化がグラフで表示される．方法。
【請求項２５】　請求項２３記載の方法であって．更に．
請求項２３の前記入力工程と前記比較工程に従って、前記基準の遺伝子発現レベルを生成する工程、を備える方法。
【請求項２６】　請求項２３記載の方法であって．更に．
前記検体の核酸配列とハイブリッドを形成する完全対合および不対合プローブのハイブリッド形成強度と．基準配列とハイブリッドを形成する完全対合および不対合プローブのハイブリッド形成強度とを．差分の閾値に対して比較する工程．を備える方法。
【請求項２７】　請求項２３記載の方法であって．更に．
前記検体の核酸配列とハイブリッドを形成する完全対合および不対合プローブのハイブリッド形成強度と．基準配列とハイブリッドを形成する完全対合および不対合プローブのハイブリッド形成強度とを．比の閾値に対して比較する工程．を備える方法。
【請求項２８】　請求項２３記載の方法であって．更に．
決定行列を用いて前記検体の核酸配列における前記遺伝子の発現変化を決定する工程、を備える方法。
【請求項２９】　請求項２３記載の方法であって．
前記検体の核酸配列における前記遺伝子の発現変化が、増加．増加傾向．減少、減少傾向、変化なし、から成るグループから選択される．方法。
【請求項３０】　コンピュータシステムにおいて．検体の核酸配列における遺伝子の発現変化を監視する方法であって、
前記遺伝子に対して完全に相補的である完全対合プローブと前記遺伝子に対して不対合な塩基を少なくとも一つ備える不対合プローブとの複数のプローブ対に関する複数のハイブリッド形成強度であって．前記完全対合および不対合プローブと前記検体の核酸配列との間のハイブリッド形成親和性を示すハイブリッド形成強度を入力す

る工程と、
前記検体の核酸配列の遺伝子発現レベルを生成するために、各対の完全対合プローブと不対合プローブのハイブリッド形成強度を比較する工程と、
前記遺伝子発現レベルを、基準の遺伝子発現レベルと比較することにより発現変化を求める工程と、を備える方法。
【請求項３１】　請求項３０記載の方法であって、更に、
請求項３０の前記入力工程と前記比較工程に従って、前記基準の遺伝子発現レベルを生成する工程、を備える方法。
【請求項３２】　請求項３０記載の方法であって、更に、
前記検体の核酸配列とハイブリッドを形成する完全対合および不対合プローブのハイブリッド形成強度と、基準配列とハイブリッドを形成する完全対合および不対合プローブのハイブリッド形成強度とを、差分の閾値に対して比較する工程、を備える方法。
【請求項３３】　請求項３０記載の方法であって、更に、
前記検体の核酸配列とハイブリッドを形成する完全対合および不対合プローブのハイブリッド形成強度と、基準配列とハイブリッドを形成する完全対合および不対合プローブのハイブリッド形成強度とを、比の閾値に対して比較する工程、を備える方法。
【請求項３４】　請求項３０記載の方法であって、更に、
決定行列を用いて、前記検体の核酸配列における前記遺伝子の発現変化を決定する工程、を備える方法。
【請求項３５】　請求項３０記載の方法であって、
前記検体の核酸配列における前記遺伝子の発現変化が、増加、増加傾向、減少、減少傾向、変化なし、から成るグループから選択される、方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、コンピュータシステムの分野に関する。更に詳しくは、本発明は、核酸配列等の生物学的配列を解析するためのコンピュータシステムに関する。
【０００２】
【従来の技術】基質上で物質の配列（アレイ）を形成し、利用するための装置やコンピュータシステムが知られている。例えば、本明細書に組み込まれるＰＣＴ出願ＷＯ９２／１０５８８には、核酸等の物質の配列を決定し、あるいは、その配列をチェックする手法が述べられている。このような操作を実行するための配列（アレイ）は、本明細書に組み込まれる米国特許第５，１４３，８５４号および米国特許出願第０８／２４９，１８８号に開示されているバイオニア的手法等の方法に従って形成される。
【０００３】これらにおいて説明されている手法に従って、核酸プローブの配列（アレイ）を、基質あるいはチップ上の既知の位置に形成する。次に、蛍光標識された核酸をチップに接触させて、標識された核酸がチップに結合している位置を示すイメージファイル（イメージファイルは処理されてセルファイルになる）をスキャナーで生成する。このセルファイルと所定の位置におけるプローブの強度とに基づいて、ＤＮＡあるいはＲＮＡのモノマー配列のような情報を得ることが可能になる。このようなシステムを用いることにより、例えば、嚢胞性線維症に関係する突然変異、（所定の癌に関係する）Ｐ５３遺伝子、ＨＩＶ等の遺伝子特性の研究および検出に利用可能なＤＮＡの配列（アレイ）が形成される。
【０００４】
【発明が解決しようとする課題】本明細書に組み込まれる米国特許出願第０８／５３１，１３７号（代理人事件番号１６５２８Ｘ－００８２１０）、第０８／５２８，６５６号（代理人事件番号１６５２８Ｘ－０１７６００）、および、第０８／６１８，８３４号（代理人事件番号１６５２８－０１６４００）に、塩基呼び出しを行うための革新的なコンピュータ利用技術が開示されている。しかし、これらのパイオニア的手法によって現在利用され、また、利用可能になっている膨大な量の情報を評価・解析処理するためには、更に、コンピュータシステムおよび方法を改良することが必要となる。
【０００５】更に、遺伝子発現を監視するためのコンピュータ利用技術の改良も必要である。多くの病態は、遺伝子ＤＮＡの複製回数の変化あるいは所定の遺伝子の（例えば、開始、ＲＮＡ前駆体の供給、ＲＮＡプロセッシング等の制御による）転写レベルの変化に起因する種々の遺伝子の発現レベルにおける相違によって特徴づけることができる。例えば、遺伝物質の損失および獲得は、悪性の形態変換及び進行に重要な役割を果たす。また、（例えば、腫瘍遺伝子あるいは腫瘍抑制遺伝子等の）所定の遺伝子の発現（転写）レベルの変化は、種々の癌の存在および進行を示す指標として働く。
【０００６】病気や細胞周期及び細胞発育の制御も、所定の遺伝子の転写レベルの変化によって特徴づけられる。例えば、ウィルス感染は、所定のウィルスの遺伝子の発現増強によって特徴づけられることが多い。例えば、単純ヘルペスウィルス感染、（伝染性単核症等の）エプスタイン－バールウィルス感染、サイトメガロウィルス感染、水痘－帯状ヘルペスウィルス感染、パルボウィルス感染、ヒトパピローマウィルス感染等の発生は、すべて、それぞれのウィルスに存在する種々の遺伝子の発現増強によって特徴づけられる。特徴的なウィルス遺伝子の増強発現レベルを検出することによって、その病態を効果的に診断することが可能になる。特に、単純ヘルペスウィルス等のウィルスは、長い潜伏期間を経て、

突然．短い時間で、爆発的に複製される。特徴的なウィルス遺伝子の発現レベルを検出することによって．このような活動的な増殖（おそらくは、その結果としての感染）状態を検出することが可能になる。

【０００７】

【課題を解決するための手段およびその作用・効果】本発明は、核酸配列等の生物学的配列（生体配列）を解析するための革新的なシステムおよび方法を提供する。コンピュータシステムは、検体配列における塩基を呼び出すために、核酸プローブと検体の核酸配列とのハイブリッド形成親和性を示すハイブリッド形成強度を解析することができる。複数の塩基コールは．単一の塩基コールを形成するために組み合わせるようにしてもよい。更に．コンピュータシステムは、遺伝子発現、あるいは、基準（基線）と比較された遺伝子発現の変化を監視するために、ハイブリッド形成強度を解析するようにしてもよい。

【０００８】本発明の一態様によれば．検体の核酸配列における未知の塩基を呼び出すためのコンピュータによる方法は、複数組の核酸プローブに関するハイブリッド形成強度であって、それぞれ核酸プローブと前記検体の核酸配列との間のハイブリッド形成親和性を示すハイブリッド形成強度を受け取る工程と、各組のプローブに関して前記未知の塩基に対する塩基コールを演算する工程と．前記複数組のプローブに関して最も発生頻度の高い前記未知の塩基に対する塩基コールに従って、前記複数組のプローブに関する単一の塩基コールを演算する工程と．を備える。典型的には．この単一塩基コールがスクリーンディスプレイ上に表示され、ユーザは、この単一塩基コールの由来である複数の塩基コールを表示するかあるいは表示しないかを選択できる。

【０００９】本発明の別の態様によれば．コンピュータによる塩基呼び出し処理に関するパラメータを動的に変更する方法は．ユーザによって変更可能なパラメータを含む前記塩基呼び出し処理を利用して．検体の核酸配列の少なくとも一部に関する複数の塩基コールを生成する工程と、前記検体の核酸配列の少なくとも一部に関する前記複数の塩基コールを表示する工程と、前記塩基呼び出し処理のパラメータを表示する工程と、前記塩基呼び出し処理のパラメータに対する新しい値を特定する入力をユーザーから受け取る工程と、前記塩基呼び出し処理と前記パラメータに対する新しい値とを用いて．前記検体の核酸配列の少なくとも一部に対する更新された塩基コールを生成する工程と．前記検体の核酸配列の少なくとも一部に対する前記更新された塩基コールを表示する工程と、を備える。典型的には、ユーザーによって変更可能なパラメータは、定数．閾値、または、範囲である。

【００１０】本発明の更に別の態様によれば、検体の核酸配列において遺伝子の発現を監視するコンピュータによる方法は、前記遺伝子に対して完全に相補的である完全対合プローブ（パーフェクトマッチプローブ）と前記遺伝子に対して不対合な塩基を少なくとも一つ備える不対合プローブ（ミスマッチプローブ）との複数のプローブ対に関する複数のハイブリッド形成強度であって、前記完全対合および不対合プローブと前記検体の核酸配列との間のハイブリッド形成親和性を示すハイブリッド形成強度を入力する工程と．各対の完全対合プローブと不対合プローブのハイブリッド形成強度を比較する工程と．前記検体の核酸配列の遺伝子発現コールを生成する工程と、を備える。好ましい実施例においては．発現コールは、発現している、どちらともいえない、または、発現していない．として示される。

【００１１】本発明の別の態様によれば、検体の核酸配列において遺伝子の発現変化を監視するコンピュータによる方法は、前記遺伝子に対して完全に相補的である完全対合プローブと前記遺伝子に対して不対合な塩基を少なくとも一つ備える不対合プローブとの複数のプローブ対に関する複数のハイブリッド形成強度であって．前記完全対合および不対合プローブと前記検体の核酸配列との間のハイブリッド形成親和性を示すハイブリッド形成強度を入力する工程と、前記検体の核酸配列の遺伝子発現レベルを生成するために．各対の完全対合プローブと不対合プローブのハイブリッド形成強度を比較する工程と．前記遺伝子発現レベルを、基準の遺伝子発現レベルと比較することにより発現変化を決定する工程と．を備える。発現の変化を、ディスプレイスクリーン上にグラフとして表示するようにしてもよい。

【００１２】本発明の性質及び利点を更に理解するために．以下に、図面に基づいて、本発明を詳細に説明する。

【００１３】

【発明の実施の形態】

概要

本発明は、検体の核酸配列においてヌクレオチドを同定し（即ち、塩基を呼び出し）、遺伝子発現を監視するための革新的な方法を提供する。以下に．本発明の好適な実施例を説明する。但し．以下の実施例は、例示に過ぎず．発明の範囲を限定するものではない。

【００１４】図１は、本発明のソフトウェアを実行するために用いられるコンピュータシステムの例を示す。図１に示すコンピュータシステム１は．モニタ３．スクリーン５、キャビネット７．キーボード９、および．マウス１１を備える。マウス１１は、マウスボタン１３等の一つあるいは複数のボタンを有する。キャビネット７には．本発明のコンピュータコードを含むソフトウェアプログラムを記憶して検索するために用いられるＣＤ－ＲＯＭドライブ１５及び（図示しない）ハードドライブが収容されている。この実施例では、ＣＤ－ＲＯＭ１７がコンピュータ読み取り可能な媒体として示されている

が、フロッピーディスク、ＤＲＡＭ、ハードドライブ、フラッシュメモリ、テープ等、他のコンピュータ読み取り可能な媒体を用いることもできる。キャビネット７には、更に、プロセッサ、メモリ等の（図示しない）周知のコンピュータ構成部品が収容されている。

【００１５】図２は、本発明のソフトウェアを実現するために用いられるコンピュータシステム１を示すシステムブロック図である。図１に示すように、コンピュータシステム１は、モニタ３とキーボード９を備える。コンピュータシステム１は、更に、中央処理装置５０、システムメモリ５２、入出力制御装置５４、ディスプレイアダプタ５６、リムーバブルディスク５８、固定ディスク６０、ネットワークインターフェース６２、スピーカ６４等のサブシステムを備える。リムーバブルディスク５８は、フロッピー、テープ、ＣＤ－ＲＯＭ、可動ハードドライブ、フラッシュメモリ等の取り外し可能なコンピュータ読み取り可能な媒体を示す。また、固定ディスク６０は、内蔵ハードドライブ等を示す。更に、コンピュータシステムに、本発明の実現に適した他のサブシステムを備えるようにしてもよい。例えば、コンピュータシステムに２つ以上の処理装置５０を備えるようにしてもよいし（即ち、マルチプロセッサシステム）、メモリキャッシュを備えるようにしてもよい。

【００１６】６６等の矢印は、コンピュータシステム１のシステムバス・アーキテクチャを示す。但し、この矢印は、サブシステムを連結するあらゆる結合スキームを示すものである。例えば、ディスプレイアダプタ５６がローカルバスを介して中央処理装置５０に接続されるようにすることもできるし、あるいは、システムがメモリキャッシュを備えるようにすることもできる。図２に示されるコンピュータシステム１は、本発明の実現に適したコンピュータシステムの一例に過ぎず、当業者に周知のように、サブシステムの他の構成において本発明を実現することも可能である。コンピュータシステムの一例として、サン・マイクロシステムズ社のワークステーションを用いることもできる。

【００１７】ＶＬＳＩＰＳ技術（商標）は、非常に小さなチップ上に、非常に大きなオリゴヌクレオチドプローブの配列（アレイ）を形成するための方法である。詳細は、本明細書に組み込まれる米国特許第５，１４３，８５４号およびＰＣＴ特許公報ＷＯ９０／１５０７０および９２／１００９２を参照のこと。チップ上のオリゴヌクレオチドプローブを用いて、対象となる検体核酸（以下、「標的」核酸）の相補的核酸配列を検出する。

【００１８】本発明は、ハイブリッド形成された核酸プローブを含むチップに関してハイブリッド形成強度ファイルを解析する方法を提供する。本実施例において、このファイルは、生物学的配列から得られる蛍光データを表しているが、放射性強度データ等の他のデータを表すようにしてもよい。即ち、本発明は、ハイブリッド形成の蛍光測定値の解析に限定されるものではなく、ハイブリッド形成の他の測定値の解析にも容易に適用可能である。

【００１９】本発明は、チップマスクを設計し、チップ上でプローブを合成し、核酸を標識して、ハイブリッド形成された核酸プローブを走査するコンピュータシステムの一部として説明する。このようなシステムは、本明細書に組み込まれる米国特許出願第０８／２４９，１８８号に詳細に説明されている。但し、本発明は、例えば遠隔位置に置かれたシステムによって生成されるデータを解析するために、全体システムとは独立に利用することも可能である。

【００２０】図３は、ＲＮＡやＤＮＡ等の生体物質の配列（アレイ）を形成し、解析するコンピュータ化されたシステムを示す。コンピュータ１００は、ＲＮＡあるいはＤＮＡ等の生体高分子の配列（アレイ）を設計するために用いられる。ここで、コンピュータ１００は、図１及び図２に示すように、例えば、適当なメモリとＣＰＵを備え、ウィンドウズＮＴ環境下にある適当にプログラムされたＩＢＭパーソナルコンピュータ互換機である。コンピュータシステム１００には、対象となる遺伝子の特性に関するユーザーからの情報、および、その配列（アレイ）の所望の特性に関する情報が入力される。あるいは、ＧｅｎＢａｎｋ等の外部あるいは内部データベース１０２から、対象となる所定の遺伝子配列に関する情報を入力するようにしてもよい。コンピュータシステム１００は、ＰＣＴ出願ＷＯ９２／１００９２に記載されているような転換行列の形で一組のチップ設計コンピュータファイル１０４および他の関連コンピュータファイルを出力する。

【００２１】チップ設計ファイルは、ＤＮＡ等の分子配列の形成に用いられるリソグラフィックマスクを設計するためのシステム１０６に与えられる。システムあるいはプロセス１０６が、マスク１１０の製造に必要なハードウェア、および、効率よくマスク上にマスクパターンを配置するために必要なコンピュータハードウェア並びにソフトウェア１０８を備えるようにしてもよい。図３の他の特性と同様に、このような構成要素を物理的に同一の位置に配置させてもさせなくてもよいが、図示の便宜上、図３では同じところに表示している。システム１０６は、マスク１１０、あるいは、クロムーガラスマスク等、高分子配列（アレイ）の形成に用いられる他の合成パターンを生成する。

【００２２】合成システム１１２では、マスク１１０およびシステム１００から入力されるチップ設計に関する所定の情報が用いられる。合成システム１１２は、基質あるいはチップ１１４上に高分子配列（アレイ）を形成するために必要なハードウェアおよびソフトウェアを備える。例えば、合成装置１１２は、光源１１６および基質あるいはチップ１１４が載置される化学フローセル１

１８を備える。マスク１１０を光源と基質／チップとの間に置いて、チップの所定領域から保護剤を除去するために、適当な回数、基質とマスクとの間を相対的に並進させる。所定の薬剤をフローセル１１８から流して、保護除去された領域への結合、および、洗浄等の操作を行う。このような操作は、全て、適当にプログラムされたコンピュータ１１９によって実行される構成が望ましい。ここで、コンピュータ１１９は、マスク設計及びマスク形成に用いられたコンピュータと同じコンピュータでも違うものでもよい。

【００２３】合成システム１１２によって形成された基質をより小さなチップに分割して、マークされた標的に曝すようにしてもよい。ここで、標的は、基質上の一つあるいは複数の分子に対して相補的であってもよく、あるいはそうでなくてもよい。また、標的は、（図３に＊印で示す）蛍光標識等の標識でマークされ、走査システム１２０内に配置される。走査システム１２０は、適当にプログラムされたデジタルコンピュータ１２２によって制御される。このデジタルコンピュータ１２２も、合成、マスク形成、およびマスク設計に用いられたコンピュータと同じコンピュータでも違うものでもよい。スキャナ１２０は、標識された標的（＊）が基質に結合した位置を検出するために用いられる共焦頭微鏡あるいはＣＣＤ（高高結合素子）等の検出装置１２４を備える。スキャナ１２０は、蛍光標識された標的の場合には、基質上の位置の関数として蛍光強度（フォトン数あるいは高圧のような他の関係する測定値）を示すイメージファイル１２４を出力する。標識された標的が高分子配列（アレイ）とより強固に結合している部分ではフォトン数が大きくなり、また、基質上の高分子中のモノマー配列は位置の関数として既知であるため、標的と相補的な基質上の高分子の配列を決定することができる。

【００２４】イメージファイル１２４は、本発明の視覚化および解析方法を組み込んだ解析システム１２６に入力される。この解析システムも種々のコンピュータシステムのいずれでもよい。本発明は、チップ設計ファイルおよびイメージファイルを解析して、適当な出力１２８を与える様々な方法を提供する。また、本発明を用いて、ＤＮＡあるいはＲＮＡ等の標的における所定の突然変異を同定するようにしてもよい。

【００２５】図４は、本発明の一実施例で用いられる全体的なソフトウェアシステムの概要を示す。図４に示されるように、システムは、まず、ステップ２０２で、所定の解析の対象となる遺伝子配列、即ち、標的を同定する。対象となる配列は、例えば、遺伝子の正常部分あるいは突然変異部分、遺伝を決定づけるあるいは修辞的情報を与える遺伝子等である。テキストファイルの手動入力によって配列を選択してもよいし、あるいは、ＧｅｎＢａｎｋ等の外部資源から選択するようにしてもよい。ステップ２０４では、システムが遺伝子の評価を行うことにより、チップ上でどのプローブが望ましいかを決定し、あるいは、ユーザーが決定するための情報を与え、チップ上でのプローブの適当な「レイアウト」を求める。

【００２６】チップは、通常、既知の配列を有する標準核酸配列に対して相補的なプローブを備える。野生型プローブは、参照配列（標準配列）と理想的にハイブリッド形成するプローブであり、野生型遺伝子（チップ野生型とも称される）は、従って、チップ上の野生型プローブと理想的にハイブリッド形成する。標的配列は、突然変異、挿入、欠失等の存在を除いては、参照配列とほぼ同じである。このレイアウトは、遺伝配列の「読み込み」及び／あるいは周縁効果の最小化、合成の容易さ等を可能にするチップ上の構成を含む所望の特性を実現するものである。

【００２７】図５は、チップの全体的なレイアウトを示す。チップ１１４は、複数のユニットから成り、各ユニットには、一つの野生型配列あるいは複数の野生型配列が様々な配置で並べられている。図にはユニット１の詳細を示してあるが、このように、各ユニットは、プローブを含むチップ上の領域である複数のセルから形成されている。各ユニットは、複数組の開追するセルを含む。ここで、セルは、一つあるいは複数の分子（例えば、核酸プローブ）の多くの複製を含む基質上の領域を意味する。

【００２８】各ユニットは、縦横（横は「レーン」）に配置された複数のセルから成る。例えば、５つの関連セルから成る組には、野生型セル２２０、「突然変異」セル２２２、「ブランク」セル２２４が含まれる。セル２２０は、野生型配列の一部の鎖体である野生型プローブを含む。セル２２２は、野生型配列に対する「突然変異」プローブを含む。例えば、野生型プローブが３’－ＡＣＧＴである場合、３’－ＡＣＡＴ、３’－ＡＣＣＴ、３’－ＡＣＧＴ、および、３’－ＡＣＴＴは「突然変異」プローブである。セル２２４は、プローブを含まない「ブランク」セルである（「ブランク」プローブとも称される）。ブランクセルにはプローブが含まれないため、標識された標的はこの領域でチップに結合しない。このため、ブランクセルの領域を用いて、バックグラウンド強度の測定が行われる。

【００２９】図４に戻って、ステップ２０６で、合成用のマスクが設計される。ステップ２０８では、ソフトウェアがマスク設計およびレイアウト情報を利用することにより、ＤＮＡ等の高分子チップが形成される。このソフトウェア２０８は、基質とマスクとの相対的な並進、所望の薬剤のフローセル内における流れ、フローセル内の合成温度等のパラメータを制御する。ステップ２１０では、他のソフトウェアを用いて、以上のようにして合成され、標識された標的に曝されたチップを走査する。ソフトウェアは、チップの走査を制御して、後に配列情

報を得るために用いられるファイルに得られたデータを格納する。

【００３０】ステップ２１２で、コンピュータシステムは、レイアウト情報と蛍光情報とを用いて、チップ上でハイブリッド形成された核酸プローブの評価を行う。ＤＮＡチップから得られる重要な情報としては、突然変異標的の同定、及び、特定の標的の遺伝子配列の決定が挙げられる。

【００３１】図６は、特定の標的ＤＮＡがＤＮＡプローブ１１４の配列（アレイ）に結合している状態を示す。この簡単な例に示されるように、配列中に以下のプローブが形成される（野生型プローブに関しては一つのプローブだけが示されている）。

【００３２】

3'－AGAACGT
AGACCGT
AGAGCGT
AGATCGT
・・・

【００３３】これらのプローブでは、それぞれ、一つの塩基のみが異なっている。即ち、調査位置において単一の塩基不対合が存在する。従って、核酸配列中のこの調査位置における塩基を同定するように、プローブが設計される。即ち、ここでは、１つのユニットとは、複数組の関連するプローブのことを言い、各組には、調査位置における単一の塩基不対合が異なる複数のプローブが含まれる。

【００３４】５'－ＴＣＴＴＧＣＡ の配列を有する蛍光標識された（あるいは別の方法でマークされた）標的をＤＮＡプローブの配列（アレイ）に曝した場合、これは、3'－AGAACGT プローブのみに相補的であり、フルオレセインは、3'－AGAACGTが位置するチップ表面上で主に観察される。一つの塩基のみがそれぞれ異なる各組のプローブに関するイメージファイルには、各々プローブに対応する４つの蛍光強度が含まれる。即ち、蛍光強度は、各々、他のプローブとは異なる各プローブのヌクレオチドあるいは塩基に結びつけられる。更に、イメージファイルには、バックグラウンドの蛍光強度として用いられる「ブランク」セルが含まれる。特定の塩基位置に関する５つの蛍光強度を解析することにより、ここに開示される本発明の方法を用いて、このような配列（アレイ）から配列（シーケンス）情報を得ることができる。

【００３５】図７は、チップ上に列で配置されたプローブを示す。参照配列（即ち野生型チップの配列）を下つきの数字で示された５つの調査位置と共に示す。調査位置は、多くの場合、標的配列が突然変異を含む、即ち、標的配列が参照配列と異なっているような、参照配列中の塩基位置である。チップには、それぞれの調査位置に対応する５つのプローブセルが含まれる。プローブセルには、各々、調査位置に共通の塩基を有するプローブの組が含まれる。例えば、参照配列は、第一の調査位置 I_1 に塩基Ｔを有する。この調査位置に関する野生型のプローブは、３'－ＴＧＡＣであり、プローブ中の塩基Ａは、参照配列における調査位置の塩基に対して相補的である。

【００３６】この第一の調査位置 I_1 に関しては、４つの「突然変異」プローブセルが存在する。４つの突然変異プローブは、それぞれ、３'－ＴＧＡＣ、３'－ＴＧＣＣ、３'－ＴＧＧＣ、３'－ＴＧＴＣ である。４つの突然変異プローブは、それぞれ調査位置において一つの塩基のみが異なっている。図示するように、野生型のプローブと突然変異プローブはチップ上に列配置されている。突然変異プローブのうちの一つ（この場合には、３'－ＴＧＡＣ）は、野生型のプローブと等しいため、突然変異を証明するものとはならない。但し、図８に示すように、冗長性により突然変異を視覚的に示すことができる。

【００３７】図７に示すように、チップは、他の調査位置 I_2－I_5 の各々に関しても、野生型プローブと突然変異プローブを有している。それぞれの場合において、野生型プローブは突然変異プローブの一つと等価である。

【００３８】図８は、チップ上の標的と図７の参照配列とのハイブリッド形成パターンを示している。比較のために、チップの一番上に参照配列を示している。チップには、ＷＴ列（野生型）、Ａ列、Ｃ列、Ｇ列、Ｔ（またはＵ）列が含まれる。各列は、プローブを含むセル列である。ＷＴ列のセルは、参照配列に相補的なプローブを含む。Ａ列、Ｃ列、Ｇ列、Ｔ列のセルは、それぞれ、列の名称となっている塩基が調査位置に存在することを除けば参照配列に対して相補的であるプローブを含んでいる。

【００３９】一実施例においては、セル内のプローブのハイブリッド形成は、標識された標的配列の結合に起因するセルの蛍光強度（例えば、フォトン数）によって決定される。蛍光強度はセルによって大きく異なる。単純化するために、図８では、領域を塗りつぶすことによって、セルが高度のハイブリッド形成をしていることを示す。ＷＴ列を観察すると、調査位置Ｉ１において野生型セルの領域が塗りつぶされていないため、この位置 I_1 に突然変異が存在することが容易にわかる。Ｃ列のセルは塗りつぶされているため、ＴからＧへの突然変異であることがわかる（突然変異プローブは相補的であるため、ＣセルはＧ突然変異を示す）。好ましい実施例では、ＷＴ列は利用されず、調査位置の塩基を呼び出すために、（「ブランク」セルを含まない）４つのセルが使用される。

【００４０】実際には、突然変異が存在する調査位置近傍のセルの蛍光強度は相対的に低く、突然変異の周囲は

「暗い領域」となる。このように蛍光強度が低くなるのは、突然変異近辺の調査位置のセルは、標的配列と完全に相補的なプローブを含んでおらず、その結果、これらのプローブと標的配列とのハイブリッド形成率が低いからである。例えば、調査位置I_2及びI_5において、それらの位置に存在するプローブはどれも標的配列に相補的でないため、セルの相対的強度が比較的低くなる。蛍光強度が低いとデータの解像度も低くなるが、本発明の方法を用いることによって、突然変異周囲の暗い領域でも精度よく塩基を呼び出すことができ、暗い領域内に存在する他の突然変異を同定することができる。

【００４１】図９は、チップ上における標準タイル構造及び代替タイル構造を示す。図示するように、チップには１２のユニット（ユニット$_{1,12}$）が含まれている。ユニット$_{1,1}$は、同一の参照配列に相補的なプローブを含むようにタイル状に構成されている（即ち、チップ上で設計され、合成されている）。以下、このユニットグループを標準グループと称する。本発明では、他に明記しない限り、標的配列に関する塩基コールには標準グループが用いられる。

【００４２】ユニット$_{1,2}$は、標準グループの参照配列とは異なった同一の参照配列に相補的であるプローブを含むようにタイル状に構成される。以下、このユニットグループを代替グループと称する。ユニット$_{2,12}$は、標準グループ及び第一の代替グループの参照配列とは異なった参照配列に基づく別の代替グループを構成している。これらの参照配列は、互いに異なってはいるが、多くの場合、かなり似ている。例えば、参照配列を、少しずつ異なったＨＩＶの突然変異としてもよい。本発明の実施例は、標的配列の塩基コールには通常用いられないような参照配列に基づくタイル構成からの情報を評価し、利用する。

【００４３】あるグループ内の複数のユニットは同一のプローブを備えるようにしてもよいし、異なった構造のプローブを備えるようにしてもよいし、あるいは、一つまたは複数の別のチップに由来するプローブを備えるようにしてもよい。例えば、１つのユニットが、プローブの第三位置に調査位置を有する５量体プローブを備えるようにしてもよい。また、別のユニットが、第六位置に調査位置を有する１０量体プローブを備えるようにしてもよい。更に、これらのユニットを同一のチップ上に配置してもよいし、異なったチップ上に配置してもよい。

【００４４】図９左下の拡大図に示すように、ユニットの各ブロックには、通常、Ａ、Ｃ、Ｇ、Ｔで示される４つのセルが含まれる。セル内の各プローブの調査位置にどの塩基があるかを、この塩基表示によって示している。各セルには、普通、何百、何千という同一の核酸プローブが存在する。

【００４５】好適な実施例では、セルは、参照配列に沿って、順番に互いに隣接して配置されているが、チップ上での位置さえ確定されている限り、特定の位置にセルを配置しなければならないという決まりはない。更に、実験を首尾一貫したものにするためには、単一のチップ上で異なったグループを合成することが望ましいが、本発明の方法を、異なったチップ上の異なった構成から得られるデータに適用することもできる。

【００４６】標的配列の解析

図１０は、チップから得られたハイブリッド形成強度を表示する画面である。解析が行われる場合、システムには、ハイブリッド形成されたチップを走査した画像を含むイメージファイルが入力される。ここで、イメージファイルが、蛍光強度と、標識された標的核酸配列あるいはその切片がチップに結合する位置を示すようにする構成も望ましい。

【００４７】画面表示（スクリーンディスプレイ）２６０には、周知のウィンドウ用ＧＵＩ（グラフィカルユーザーインターフェース）が用いられている。ユーザーの選択に従って、イメージファイルが表示される。ユーザーがイメージファイルの表示を選択すると、イメージファイルを含むウィンドウ２６２が表示される。図示されているイメージファイルには、Ａ列、Ｃ列、Ｇ列、Ｔ列の複数の列が含まれる。

【００４８】ユーザーが表示されているイメージファイル上にカーソルを動かすと、ステータスバー２６４に、カーソルのＸ、Ｙ座標の位置とその位置における蛍光強度が示される。また、ポインティング・デバイスを用いて、イメージファイルの矩形の領域を選択することにより、サブイメージを処理することも可能である。例えば、サブイメージを拡大して、個々のセルがもっとはっきりと見えるようにすることもできる。更に、強度のコントラストを調整することによって、現在のコントラストの設定ではばっきりしないハイブリッド形成強度の差異を明確にすることもできる。

【００４９】図１１は、関連するプローブのハイブリッド形成強度に基づいて塩基呼び出し処理を行う方法を示すフローチャートである。ここで、「関連するプローブ」とは、調査位置におけるヌクレオチド塩基が異なるプローブを示す。通常、これらのプローブは調査位置を除いて同一の構成をしているが、更に、別の塩基位置でも異なる構成にしてもよい。即ち、関連するプローブでは、少なくとも一つの塩基が異なっている。

【００５０】ステップ３０２で、バックグラウンド強度、即ち「ブランク」セル強度を差し引くことによって、４つの関連するプローブのハイブリッド形成強度を調整する。引き算の結果ハイブリッド形成強度がゼロ以下になる場合には、小さな正の数をハイブリッド形成強度に設定することが望ましい。このように設定することにより、以降の計算で、ゼロあるいは負の数での割り算を避けることができる。

【００５１】ステップ３０４では、ハイブリッド形成強

度を強度順に並べる。次に、ステップ３０６で、最大強度を、所定のバックグラウンド差分限界値と比較する。バックグラウンド差分限界値は、未知の塩基を正しく呼び出すために、最高強度を有するプローブが超えなければならないハイブリッド形成強度を示す数値である。即ち、バックグラウンド調整後の塩基強度は、バックグラウンド差分限界値より大きくなければならない。そうでない場合には、未知の塩基を正確に呼び出すことができなくなる可能性がある。

【００５２】関連するプローブで最も高いハイブリッド形成強度がバックグラウンド差分限界値以下の場合には、ステップ３０８で、（不十分な強度を示す）コードＮを未知の塩基に割り当てる。そうでない場合には、ステップ３１０で、最高のハイブリッド形成強度と二番目に高いハイブリッド形成強度との比を計算する。

【００５３】次に、ステップ３１２で、ステップ３１０で計算した比を、所定の比限界値と比較する。比限界値は、未知の塩基を同定するために必要な比を示す数値である。好ましい実施例においては、比限界値は１．２である。計算された比が比限界値より大きければ、最高のハイブリッド形成強度を有するプローブに従って、未知の塩基が呼び出される。具体的には、ステップ３１４で、最高強度のプローブ内の調査位置における塩基の補体を未知の塩基として呼び出す。一方、計算された比が比限界値以下である場合には、ステップ３１６で、二番目に高いハイブリッド形成強度と三番目に高いハイブリッド形成強度との比を計算する。

【００５４】次に、ステップ３１８で、ステップ３１６で計算した比を比限界値と比較する。計算された比が比限界値より大きい場合には、ステップ３２０で、一番目と二番目に高いハイブリッド形成強度を有するプローブの調査位置における塩基の補体を規定する不確定コードを、未知の塩基として呼び出す。一方、計算された比が比限界値以下である場合には、ステップ３２２で、三番目に高いハイブリッド形成強度と四番目に高いハイブリッド形成強度との比を計算する。

【００５５】次に、ステップ３２４で、ステップ３２２で計算した比を比限界値と比較する。計算された比が比限界値より大きい場合には、ステップ３２６で、一番目ないし三番目に高いハイブリッド形成強度を有するプローブの調査位置における塩基の補体を規定する不確定コードを、未知の塩基として呼び出す。一方、計算された比が比限界値以下である場合には、ステップ３２８で、（不十分な区別を示す）コードＸを未知の塩基に割り当てる。

【００５６】図１２は、関連するプローブのハイブリッド形成強度に基づいて塩基呼び出し処理を行う他の方法を示すフローチャートである。このフローチャートでは、関連するプローブが示すハイブリッド形成強度に基づいて処理が行われる。即ち、塩基コール処理により、調査位置における一つの塩基の不対合だけが互いに異なる複数のプローブ内の調査位置に対応するような、標的内の塩基が呼び出される。ステップ４０２で、システムは、標的配列に対して最高のハイブリッド形成強度を有するプローブが一つだけ存在するかどうかを判定する。存在しない場合には、不確定を意味するＮを塩基に割り当てる。例えば、２つのプローブが同じ最高強度を有している場合（即ち、同強度である場合）、その塩基にＮを割り当てる。

【００５７】標的に対して最高のハイブリッド形成強度を有するプローブが一つだけ存在する場合には、ステップ４０６で、そのプローブに従って塩基の呼び出しが行われる。プローブは標的配列に相補的であるため、プローブの調査位置における塩基に相補的な塩基（Ｃ／Ｇ、Ａ／Ｔ）を呼び出すようにしてもよい。

【００５８】次に、ステップ４０８で、システムは、塩基コール（呼び出された塩基）が突然変異であるかどうかを判定する。即ち、参照配列の塩基と異なっているかどうかを判定する。塩基コールが突然変異塩基コールでない場合には、塩基コールを確定する。一方、塩基コールが突然変異塩基コールの場合には、システムは、ステップ４１０で所定の「突然変異」条件を満たしているかどうかを判定し、満たしていない場合には、ステップ４１２で、Ｎを塩基に割り当てる。

【００５９】以下に突然変異条件の例を説明するが、説明をわかりやすくするために、関連するプローブのハイブリッド形成強度を次のようにラベルする。HighInt は最高のハイブリッド形成強度を、SecondInt は二番目に高いハイブリッド形成強度を、ThirdIntは三番目に高いハイブリッド形成強度を、LowIntは最も低いハイブリッド形成強度をそれぞれ示す。

【００６０】例えば、呼び出された塩基が突然変異であると規定する突然変異条件は、以下の３つのテスト全てを満足させるものでなければならない。第一のテストは、HighInt とSecondInt との差が差分限界値より大きいかどうかを判定するものである。即ち、システムは、（HighInt－SecondInt）が所定の値より大きいかどうかを判定する。この所定の値は、最高のハイブリッド形成強度が次に高いハイブリッド形成強度よりも所望の量だけ大きい場合にのみ、呼び出された塩基が突然変異塩基の可能性があると認めるように選択されたものである。

【００６１】第二のテストは、第一の比が第一の比限界値未満であるかどうかを判定するものである。第一の比は以下のように規定される。

【００６２】{SecondInt－√(ThirdInt*LowInt)}/ {HighInt-√(ThirdInt*LowInt)}

【００６３】システムは、この第一の比が所定の値未満であるかどうかを判定する。この所定の値は、２つのより低いハイブリッド形成強度を差し引いた後でも、最高のハイブリッド形成強度が、次に高いハイブリッド形成

強度に対して所望の比より大きい場合にのみ、呼び出された塩基が突然変異塩基の可能性があると認めるように選択されたものである。

【００６４】第三のテストは、隣接比が隣接比限界値より大きいかどうかを判定するものである。隣接比は以下のように規定される。

【００６５】$HighInt_n / \{HighInt_n - \sqrt{(HighInt_{n+1} * HighInt_{n-1})}\}$

【００６６】ここで、添え字のｎは、呼び出されている対象の塩基位置における値を示す。また、ｎ＋１及びｎ－１は、隣接する塩基位置における値を示す。システムは、この隣接比が所定の値より大きいかどうかを判定する。この所定の値は、最高のハイブリッド形成強度が、隣接する最高のハイブリッド形成強度を引いた後の最高ハイブリッド形成強度に対して所望の比より大きい場合にのみ、呼び出された塩基が突然変異塩基の可能性があると認めるように選択されたものである。

【００６７】このように、突然変異条件の全てが満たされた場合にのみ、呼び出された塩基が突然変異塩基と規定される。突然変異はかなりまれであるため、突然変異が存在する可能性が高い場合にのみ突然変異塩基を呼び出すようにする。突然変異条件が満たされない場合には、呼び出された塩基は不確定であるか、あるいは、標準塩基と同一である（統計的に、正しい塩基コールの可能性がある）。

【００６８】この実施例では、３つの突然変異条件を用いたが、突然変異条件を一つのみ（例えば、上述の条件のどれか一つ）とすることもできる。また、参照文献として先に挙げた米国特許出願に記載されているような塩基呼び出し方法を利用することもできる。

【００６９】図１３は、一つのユニットグループに関して塩基呼び出し処理を行う方法を示すフローチャートである。上述したように、一つのユニットには、複数組の関連セルが含まれ、関連セルは、調査位置における一つの塩基がそれぞれ異なっているプローブを含む。システムは、まず、（例えば、ハイブリッド形成されたチップを走査するスキャナによって生成されたイメージデータファイルから）ハイブリッド形成強度を入力し、更に、このハイブリッド形成強度に対応するプローブの構造を入力する。次に、バックグラウンド強度（例えば、「ブランク」セルの強度、あるいは、プローブを含まないチップの別の領域において測定された強度）を入力されたハイブリッド形成強度から引く。ここで、バックグラウンド引き算後のハイブリッド形成強度の最低値を１（フォトン数１）に設定しておいてもよい。

【００７０】ハイブリッド形成強度は、プローブ（あるいはプローブの複数の複製）と標的配列との間で測定されるハイブリッド形成の度合いを意味する。例えば、測定されたセルのフォトン数の平均をハイブリッド形成強度としてもよい。ここで、フォトン数は、セル内のプローブに結合し、フルオレセインで標識された標的配列に由来する。

【００７１】ステップ４５２で、システムは、コール処理の対象となる標的配列の塩基位置を入力する。次に、ステップ４５４で、グループの各ユニットに関してその塩基位置における塩基コール処理を実行する。具体的には、その塩基位置における各ユニットの関連セルのハイブリッド形成強度を解析する。この解析を行うことにより（解析処理の詳細に関しては、図１１及び図１２で既に説明した）、システムは、各ユニットに関する塩基コールを求める。例えば、グループ内に５つのユニットが存在する場合には、５つの塩基コールが生成される。

【００７２】システムは、ステップ４５６で、そのグループのユニットに関する複数の塩基コールを解析して、そのグループの塩基コールを一つ決定する。具体的には、そのグループのユニットから呼び出される頻度が一番高かった塩基を、そのグループの塩基として呼び出すように構成してもよい。例えば、５つのユニットが存在し、各ユニットが次のような塩基コールがなされたとする。

【００７３】「Ｔ」－３ユニット
「Ｇ」－１ユニット
「Ｎ」－１ユニット

【００７４】この場合には、５ユニットのうち３ユニットが呼び出しているＴが、塩基として呼び出される。同じ頻度で呼び出された塩基が複数ある場合には、塩基を呼び出すユニットの最高平均ハイブリッド形成強度等の他のファクターを考慮する。本発明の実施例においては、図１５に示される方法が用いられる。

【００７５】ステップ４５８で、解析を実行すべき次の塩基位置が存在するかどうかを判定する。本発明の方法を適用して、標的核酸配列の全ての塩基位置に関して塩基コール処理を実行することもできるし、不必要な塩基位置をスキップして、所定の塩基位置においてのみ塩基コール処理を実行することもできる。

【００７６】図１４は、複数のユニットグループに関して塩基呼び出し処理を行う方法を示すフローチャートである。図９に示したように、解析対象となる一つあるいは複数のチップ上に複数のグループが存在する場合がある。複数のグループは、異なった参照配列に従って配置されていてもよい。但し、これは、複数のグループのハイブリッド形成に関する情報が利用されない可能性を示すものではない。通常、標準グループの参照配列が標的配列と一致する可能性が最も高いが、代替グループの一つの方が一致の可能性が高い場合には（即ち、塩基コール処理により適している場合には）、そのグループを用いて塩基処理を行う。

【００７７】ステップ５０２で、システムは、標準グループ及び代替グループのユニットに関して塩基コール処理を実行する。塩基コール処理は、例えば、先に図１３

で説明したように行われる。

【００７８】次に、ステップ５０４で、各ユニットグループに関して塩基コールを一つ求める。ここで、ユニットからの呼び出し頻度が最も高かった塩基を塩基コールとしてもよいし、あるいは、図１５に従って後に詳細を説明する方法で、各グループの塩基コールを決定するようにしてもよい。

【００７９】各ユニットグループ（即ち、標準グループと代替グループ）に関して一つの塩基コールをもとめた後、ステップ５０６で、塩基位置が入力される。次に、入力された塩基位置に関して塩基コールを実行するために最適のユニットグループを選択する。具体的には、例えば、調査位置近傍、即ち、調査位置周囲のウィンドウにおいて、標的配列に対する不対合が最も少ないグループの参照配列を求めることにより、最適グループの選択が行われる。調査位置近傍において不対合が最も少ないユニットグループは、最も精度の高い塩基コールを生成する可能性が高い。最適グループを選択する処理に関する詳細は、図１６を参照して後述する。

【００８０】次に、ステップ５１０で、最適のユニットグループに従って、入力された塩基位置における塩基の呼び出し処理を行う（即ち、ステップ５０４で決定されたそのグループの最適コールを用いる）。塩基コールが決定されると、処理は次のステップに移り、塩基コール処理を実行すべき次の塩基位置が存在するかどうかの判定が行われる。次の塩基位置が存在する場合には、ステップ５０６に戻り、次の塩基位置に関して塩基コール処理を実行する。

【００８１】図１５は、一つのユニットグループに関して塩基呼び出し処理を行う方法を示すフローチャートである。ステップ６０２で、システムは、所定の塩基位置において大部分のユニットが同じ塩基を呼び出しているかどうかを判定する。この場合、（不確定を示すコードNが割り当てられたものを除き）塩基を呼び出すユニットのみが判定の対象となる。例えば、７つのユニットが存在し、各ユニットが次のような塩基コールを実行したとする。

【００８２】「G」－３ユニット
「T」－１ユニット
「N」－４ユニット

【００８３】この場合には、４つの確定塩基コールのうち３つがGを呼び出しているため、まず、そのユニットグループに関してGを塩基として呼び出す。例外規則が適用される場合を除いて、ステップ６０４で、この塩基が多数塩基として呼び出される。

【００８４】例外規則は、そのユニットグループに関してどの塩基を呼び出すかを決める条件を規定したものである。この例外規則は、多数塩基コールを変える条件や、大多数のユニットによって呼び出される塩基が存在しないような状況に対処する条件を規定する。例外規則の一例として、（一つのユニットがある塩基を呼び出し、他のユニットが別の塩基を呼び出すような場合に適用される）隣接するプローブのハイブリッド形成強度を解析するタイブレーク規則が挙げられる。また、別の例として、３つのユニットがそれぞれ別の塩基を呼び出し、そのうちの一つが標準塩基を呼び出すものである場合、標準塩基をそのユニットグループの塩基コールとする例外規則が挙げられる。他の例外規則に関しては、後述に示す。

【００８５】ステップ６０６で、例外規則が適用されるかどうかの判定がなされる。例外規則が適用されると判定された場合には、ステップ６０８で例外規則が適用される。

【００８６】図１６は、塩基コール処理を実行するために最適のユニットグループを選択する方法を示すフローチャートである。上述したように、調査位置近傍において、標的配列に対する不対合が最も少ないグループの参照配列を求めることにより、最適グループの選択が行われる。調査位置近傍において不対合が最も少ないユニットグループは、最も精度の高い塩基コールを生成する可能性が高い。解析の対象となる調査位置近傍のウィンドウは、所定の値に設定してもよく、あるいは、プローブ構造に従って設定してもよい。例えば、全てのグループのプローブに関して、調査位置からの最大距離が、調査位置の一方の側に対しては８塩基位置、また、もう一方の側に対しては１０塩基位置である場合、この範囲の塩基位置を含むようにウィンドウを設定してもよい。

【００８７】ステップ７０２で、システムは、標準ユニットグループおよび代替ユニットグループに関して不対合得点を計算する。不対合得点は、標的配列に対する参照配列の不対合の数を示すものである。不対合得点を求めるために、参照配列のうち少なくとも２つが異なっている塩基位置のみを解析するようにしてもよい。全ての参照配列が或る塩基位置において同一である場合には、この塩基配列をスキップすることができる。

【００８８】少なくとも２つの参照配列が異なっている各塩基位置において、あるグループに関して呼び出された塩基（標的配列における有望な塩基を示す塩基コール）が参照配列の対応する塩基と異なっているかどうかを判定する。ここで、塩基コールと参照配列の対応する塩基とが異なっている場合には、不対合得点に１が加えられる。各グループに関する不対合得点は、最初にゼロに設定される。

【００８９】不対合得点は、標的配列と異なっている参照配列部分の塩基位置の数を示すものだと考えることができる（全ての参照配列において同一である塩基位置を除くようにしてもよい）。この概念をさらに説明するために、以下に簡単な例を挙げる。この例では、以下のような１つの標準グループと２つの代替グループが存在する。

で説明したように行われる。

【００７８】次に、ステップ５０４で、各ユニットグループに関して塩基コールを一つ求める。ここで、ユニットからの呼び出し頻度が最も高かった塩基を塩基コールとしてもよいし、あるいは、図１５に従って後に詳細を説明する方法で、各グループの塩基コールを決定するようにしてもよい。

【００７９】各ユニットグループ（即ち、標準グループと代替グループ）に関して一つの塩基コールをもとめた後、ステップ５０６で、塩基位置が入力される。次に、入力された塩基位置に関して塩基コールを実行するために最適のユニットグループを選択する。具体的には、例えば、調査位置近傍、即ち、調査位置周囲のウィンドウにおいて、標的配列に対する不対合が最も少ないグループの参照配列を求めることにより、最適グループの選択が行われる。調査位置近傍において不対合が最も少ないユニットグループは、最も精度の高い塩基コールを生成する可能性が高い。最適グループを選択する処理に関する詳細は、図１６を参照して後述する。

【００８０】次に、ステップ５１０で、最適のユニットグループに従って、入力された塩基位置における塩基の呼び出し処理を行う（即ち、ステップ５０４で決定されたそのグループの最適コールを用いる）。塩基コールが決定されると、処理は次のステップに移り、塩基コール処理を実行すべき次の塩基位置が存在するかどうかの判定が行われる。次の塩基位置が存在する場合には、ステップ５０６に戻り、次の塩基位置に関して塩基コール処理を実行する。

【００８１】図１５は、一つのユニットグループに関して塩基呼び出し処理を行う方法を示すフローチャートである。ステップ６０２で、システムは、所定の塩基位置において大部分のユニットが同じ塩基を呼び出しているかどうかを判定する。この場合、（不確定を示すコードNが割り当てられたものを除き）塩基を呼び出すユニットのみが判定の対象となる。例えば、７つのユニットが存在し、各ユニットが次のような塩基コールを実行したとする。

【００８２】「G」－３ユニット
「T」－１ユニット
「N」－４ユニット

【００８３】この場合には、４つの確定塩基コールのうち３つがGを呼び出しているため、まず、そのユニットグループに関してGを塩基として呼び出す。例外規則が適用される場合を除いて、ステップ６０４で、この塩基が多数塩基として呼び出される。

【００８４】例外規則は、そのユニットグループに関してどの塩基を呼び出すかを決める条件を規定したものである。この例外規則は、多数塩基コールを変える条件や、大多数のユニットによって呼び出される塩基が存在しないような状況に対処する条件を規定する。例外規則の一例として、（一つのユニットがある塩基を呼び出し、他のユニットが別の塩基を呼び出すような場合に適用される）隣接するプローブのハイブリッド形成強度を解析するタイブレーク規則が挙げられる。また、別の例として、３つのユニットがそれぞれ別の塩基を呼び出し、そのうちの一つが標準塩基を呼び出すものである場合、標準塩基をそのユニットグループの塩基コールとする例外規則が挙げられる。他の例外規則に関しては、後述に示す。

【００８５】ステップ６０６で、例外規則が適用されるかどうかの判定がなされる。例外規則が適用されると判定された場合には、ステップ６０８で例外規則が適用される。

【００８６】図１６は、塩基コール処理を実行するために最適のユニットグループを選択する方法を示すフローチャートである。上述したように、調査位置近傍において、標的配列に対する不対合が最も少ないグループの参照配列を求めることにより、最適グループの選択が行われる。調査位置近傍において不対合が最も少ないユニットグループは、最も精度の高い塩基コールを生成する可能性が高い。解析の対象となる調査位置近傍のウィンドウは、所定の値に設定してもよく、あるいは、プローブ構造に従って設定してもよい。例えば、全てのグループのプローブに関して、調査位置からの最大距離が、調査位置の一方の側に対しては８塩基位置、また、もう一方の側に対しては１０塩基位置である場合、この範囲の塩基位置を含むようにウィンドウを設定してもよい。

【００８７】ステップ７０２で、システムは、標準ユニットグループおよび代替ユニットグループに関して不対合得点を計算する。不対合得点は、標的配列に対する参照配列の不対合の数を示すものである。不対合得点を求めるために、参照配列のうち少なくとも２つが異なっている塩基位置のみを解析するようにしてもよい。全ての参照配列が或る塩基位置において同一である場合には、この塩基配列をスキップすることができる。

【００８８】少なくとも２つの参照配列が異なっている各塩基位置において、あるグループに関して呼び出された塩基（標的配列における有望な塩基を示す塩基コール）が参照配列の対応する塩基と異なっているかどうかを判定する。ここで、塩基コールと参照配列の対応する塩基とが異なっている場合には、不対合得点に１が加えられる。各グループに関する不対合得点は、最初にゼロに設定される。

【００８９】不対合得点は、標的配列と異なっている参照配列部分の塩基位置の数を示すものだと考えることができる（全ての参照配列において同一である塩基位置を除くようにしてもよい）。この概念をさらに説明するために、以下に簡単な例を挙げる。この例では、以下のような１つの標準グループと２つの代替グループが存在する。

て．簡単に解析を行うことができるように、参照配列と検体配列とが自動的に並べられる。

【０１０１】図１７は、スクリーンディスプレイの配置を簡単に示すものであるが．図１８に示すように．チップファイル及び複合ファイルから入力された情報を出さないようにしたり（非表示にしたり）．まとめたりすることもできる。例えば、ユーザーが．スクリーン領域８０８において．複合ファイルのファイル名の前につけられているスクリーンアイコンの＋マークを「クリックする」．即ち起動した場合には、その複合ファイルに関する詳細な情報が表示される。図示されるように．チップファイルから入力された情報を組み合わせるために用いられた方法を．個々のチップファイルと共に表示するようにしてもよい。

【０１０２】また、ユーザーが、スクリーン領域８０８において、チップファイルのファイル名の前につけられているスクリーンアイコンの＋マークを起動した場合には．塩基の呼び出しに用いられた工程あるいは方法を含むチップファイルに関する詳細な情報が表示される。図１８の例では．図１０に基づいて説明した「比に基づくアルゴリズム」が塩基呼び出し処理に用いられている。さらに、ユーザーは、スクリーン上に表示されている塩基コールに直ちに反映される塩基呼び出し処理パラメータを変更することもできる。例えば．この例では．比の限界値（「Ratio」）が１．２に設定されているが、ユーザーがこの比の限界値を１．４に増加させた場合には．そのチップに関する塩基コールを再演算して．新しい塩基コールをスクリーン領域８１２に表示する。塩基コール処理パラメータは．定数、閾値．範囲等のいずれの値でもよい。

【０１０３】図１８に示すように、ユーザーが理解しやすいように、（様々な構成を含む）複数の実験から得られたデータを組み合わせることができる。図１７に示される検体配列４４０－２Ａは、図１８では拡大されて、この塩基コールが複数の実験から得られたものであることがわかる。ここで、複数の実験から得られたデータは．一つのチップに関するものでも．あるいは．複数のチップに関するものでもよい。言いかえると、図１７および図１８で検体配列４４０－２Ａとして示されるヌクレオチド配列は．一つの実験結果を示すものではなく、複数の実験結果を組み合わせた、あるいは、まとめたものである。図１８に示すように、ユーザーは、それぞれの実験から得られたデータをまとめることができる。図１８の例では．それぞれの塩基に関するハイブリッド形成強度が表示されている。塩基位置１００に示すように．解析中の塩基位置を反転表示することもできる。

【０１０４】図１７に示すように、検体配列４４０－２Ａの名称の前にもスクリーンアイコンの＋マークが表示されている。このスクリーンアイコンを起動することにより．複合塩基コールを構成する個々の塩基コールを表示することができる。また．図１８に示すように．複合塩基コールは．複数の塩基コールから誘導されている。この複数の塩基コールは．塩基位置に従って、複合塩基コールと並べられる。本発明の構成により、ユーザーは．配列を解析する際に．必要に応じて、データを表示したり、非表示にしたり．あるいは．まとめたりすることができる。

【０１０５】適正子発現の監視

図１９は、複数対の完全対合プローブと不対合プローブのハイブリッド形成強度を比較することにより適正子の発現を監視する方法を示すフローチャートである。「完全対合プローブ」は、所定の標的配列に完全に相補的な配列を有するプローブを意味する。テストプローブは、通常．標的配列の一部（サブ配列）に完全に相補的である。また、「不対合対照」あるいは「不対合プローブ」は．所定の標的配列に完全に相補的でないように故意的に選択された配列を有するプローブを意味する。高密度配列（アレイ）における不対合（ＭＭ）対照には．それぞれ．同じ所定の標的配列に対して完全に相補的な対応する完全対合（ＰＭ）プローブが存在する。

【０１０６】基質表面あるいはチップ表面に望ましくは共有結合的に結合されている、複数対の完全対合プローブ及び不対合プローブのハイブリッド形成強度が比較される。核酸プローブの密度としては．基質１ｃｍ2 当たりに約６０個以上の異なった核酸プローブが含まれる構成が最も望ましい。図１９のフローチャートでは．各ステップが順に示されているが、必ずしも、これらのステップをこの順番で実行しなくてもよい。当業者に周知のように、本発明の範囲から逸脱することなく、これらのステップの順番を変えたり．組み合わせたり、あるいはいくつかのステップを除くこともできる。

【０１０７】まず最初に．標的配列（あるいは適正子）に相補的な核酸プローブを選択する。これらのプローブは完全対合プローブである。次に、標的配列に対して完全に相補的でないように構成されたプローブを選択する。これらのプローブは不対合プローブであり．不対合プローブには．それぞれ．完全対合プローブに対して少なくとも一つのヌクレオチド不対合が含まれる。従って．不対合プローブおよびそれに対応する完全対合プローブにより、一対のプローブが形成される。上述したように．不対合プローブの中心付近にヌクレオチド不対合が存在するような構成が望ましい。

【０１０８】完全対合プローブの長さは、通常．標的配列に対する高いハイブリッド形成親和性を示すように選択される。例えば、核酸配列を全て２０量体としてもよい。また、不確定さを排除する等様々な理由により、基質上で、異なった長さのプローブを合成するようにしてもよい。

【０１０９】上述したように、標的配列は、分割され、標識されて、核酸プローブを含む基質に曝される。核酸

プローブのハイブリッド形成強度を測定して、コンピュータシステムに入力する。ここで用いられるコンピュータシステムは、基質のハイブリッド形成を指示するシステムと同じものでも異なったものでもよい。いずれのコンピュータシステムを用いるにしても、遺伝子名、遺伝子配列、プローブ配列、基質上でのプローブ位置等、他の実験の詳細もシステムに入力される。

【０１１０】図１９に示すように、ハイブリッド形成後、ステップ９０２で、コンピュータシステムに、複数対の完全対合プローブと不対合プローブのハイブリッド形成強度が入力される。ハイブリッド形成強度は、核酸プローブと（遺伝子に対応する）標的核酸との間のハイブリッド形成親和性を示すものである。各対には、標的核酸の一部に完全に相補的な完全対合プローブと、少なくとも一つのヌクレオチドが完全対合プローブと異なっている不対合プローブとが含まれる。

【０１１１】ステップ９０４で、コンピュータシステムが、各対の完全対合プローブと不対合プローブのハイブリッド形成強度を比較する。遺伝子が発現している場合には、完全対合プローブのハイブリッド形成強度（即ち親和性）は、対応する不対合プローブの強度よりも有意に高い。一般に、対のプローブのハイブリッド形成強度がほとんど同じであれば、遺伝子が発現されていないと考えられる。但し、一対のプローブに基づいて判定を行うわけではなく、多数対のプローブの解析に基づいて、遺伝子が発現しているかどうかの判定が行われる。複数対のプローブのハイブリッド形成強度を比較する具体的な方法の詳細に関しては、図２０を参照して説明する。

【０１１２】完全対合プローブと不対合プローブのハイブリッド形成強度を比較した後、ステップ９０６で、遺伝子の発現状態を表示する。例えば、遺伝子が存在している（発現している）、どちらともいえない、あるいは、存在していない（発現していない）ことを示す発現コールをユーザーに表示するようにしてもよい。

【０１１３】図２０は、決定行列を用いて遺伝子が発現しているかどうかを判定する方法を示すフローチャートである。ステップ９５２で、コンピュータシステムに、Ｎ対の完全対合プローブと不対合プローブに関する生の走査データを入力する。好適な実施例において、ハイブリッド形成強度は、基質上でプローブにハイブリッド結合されるフルオレセイン標識標的から得られるフォトン数である。以下、完全対合プローブのハイブリッド形成強度をi_{pm}で示し、不対合プローブのハイブリッド形成強度をi_{mm}で示す。

【０１１４】ステップ９５４で、一対のプローブのハイブリッド形成強度を検索する。次のステップ９５６で、バックグラウンド信号強度を、その対の各ハイブリッド形成強度から差し引く。全ての生走査データから同時にバックグラウンドを差し引くようにしてもよい。

【０１１５】ステップ９５８で、プローブ対のハイブリッド形成強度を差分閾値（Ｄ）および比の閾値（Ｒ）と比較する。即ち、プローブ対のハイブリッド形成強度の差（$i_{pm}-i_{mm}$）が差分閾値以上であるかどうか、且つ、プローブ対のハイブリッド形成強度の比（i_{pm}/i_{mm}）が比の閾値以上であるかどうかが判定される。これらの閾値は、通常、一つあるいは複数の遺伝子の発現状態を精度よく監視できるようにユーザーが定義した値である。例えば、差分閾値を２０に比の閾値を１．２に設定するようにしてもよい。

【０１１６】$I_{pm}-I_{mm}>=D$、及び、$I_{pn}/I_{mn}>=R$の場合には、ステップ９６０でＮＰＯＳを１増加させる。ＮＰＯＳは、遺伝子が発現する可能性が高いことを示すハイブリッド形成強度を有するプローブ対の数を示す値である。即ち、遺伝子の発現状態の決定に、ＮＰＯＳが用いられる。

【０１１７】ステップ９６２では、$I_{mm}-I_{pn}>=D$及び$I_{mm}/I_{pn}>=R$が成立するかどうかの判定がなされる。この式が成り立つ場合には、ステップ９６４でＮＮＥＧを１増加させる。ＮＮＥＧは、遺伝子が発現しない可能性が高いことを示すハイブリッド形成強度を有するプローブ対の数を示す値である。ＮＰＯＳと同様に、ＮＮＥＧも遺伝子の発現状態の決定に用いられる。

【０１１８】遺伝子が発現していることを示す、あるいは、発現していないことを示すハイブリッド形成強度を有する各プローブ対に関して、ステップ９６６で、比の対数値（ＬＲ）と強度差値（ＩＤＩＦ）を計算する。ＬＲは、プローブ対のハイブリッド形成強度の比（i_{pm}/i_{mm}）の対数をとることにより求められる。また、ＩＤＩＦは、プローブ対のハイブリッド形成強度の差（$i_{pm}-i_{mm}$）をとることにより求められる。ステップ９６８で次のプローブ対のハイブリッド形成強度が存在すると判定されれば、処理をステップ９５４に戻して、検索を行う。

【０１１９】ステップ９７２で、決定行列を用いて、遺伝子が発現しているかどうかを判定する。決定行列では、Ｎ、ＮＰＯＳ、ＮＮＥＧ、ＬＲ（複数のＬＲ）の値が用いられ、次の計算が行われる。

【０１２０】Ｐ１　＝　ＮＰＯＳ／ＮＮＥＧ
Ｐ２　＝　ＮＰＯＳ／Ｎ
Ｐ３　＝　（１０＊ＳＵＭ（ＬＲ））／（ＮＰＯＳ＋ＮＮＥＧ）

【０１２１】これらのＰ値を用いて、遺伝子が発現しているかどうかを判定する。

【０１２２】説明のために、Ｐ値をいくつかの範囲に分割する。Ｐ１が２．１以上であれば、Ａが真である。Ｐ１が２．１未満で１．８以上であれば、Ｂが真である。それ以外の場合には、Ｃが真である。即ち、Ｐ１は、Ａ、Ｂ、Ｃの３つの範囲に分割される。これは、発明の理解を深めることを目的としてなされるものである。このようにして、全てのＰ値を以下のようにいくつかの範

29

囲に分割する。

【０１２３】Ａ＝（Ｐ１＞＝２．１）
Ｂ＝（２．１＞Ｐ１＞＝１．８）
Ｃ＝（Ｐ１＜１．８）

【０１２４】Ｘ＝（Ｐ２＞＝０．３５）
Ｙ＝（０．３５＞Ｐ２＞＝０．２０）
Ｚ＝（Ｐ２＜０．２０）

【０１２５】Ｑ＝（Ｐ３＞＝１．５）
Ｒ＝（１．５＞Ｐ３＞＝１．１）
Ｓ＝（Ｐ３＜１．１）

【０１２６】上記のブール値に従ってＰ値を所定の範囲に分割した後、遺伝子発現の判定を行う。

存在	A and (X or Y) and (Q or R)
	B and X and Q

【０１３０】

どちらともいえない	A and X and S
	B and X and R
	B and Y and (Q or R)

【０１３１】

不存在　　それ以外の場合（例えば、Ｃの組み合わせ）

【０１３２】ステップ９７４で、ユーザーに結果が表示される。例えば、存在をＰ、どちらともいえないをＭ、不存在をＡのように出力するようにしてもよい。

【０１３３】全てのプローブ対を処理して、遺伝子の発現状態を表示した後、ステップ９７５で、ＬＲを１０倍した値の平均値を計算する。さらに、ＮＰＯＳ及びＮＮＥＧを増加させたプローブに関するＩＤＩＦ値の平均を計算する。この平均値を、発現レベルを示す値として用いることもできる。また、これらの値を用いて、この実験と他の実験の定量的な比較を行うようにしてもよい。

【０１３４】ステップ９７６で定量的な測定を行う。一例として、現在の実験データを前の実験データと比較する（この場合、例えば、ステップ９７０で計算された値を用いる）。別の例として、生物検体中に存在する（バクテリア由来等の）既知量のＲＮＡのハイブリッド形成強度と実験データを比較することもできる。このようにして、遺伝子発現状態の表示即ちコールが正しいかどうかを証明することもできるし、閾値を修正することもできるし、さらに、先に行われた実験の修正を行うことも可能である。

【０１３５】説明を簡単にするために、図２０では一つの遺伝子に関する処理だけを説明したが、生物検体に含まれる複数の遺伝子の処理にこの方法を適用することもできる。従って、一つの遺伝子の解析に関する説明は、その方法を複数の遺伝子の処理に拡張できないことを示すものではない。

【０１３６】図２１は、遺伝子発現監視ソフトウェアの画面表示（スクリーンディスプレイ）のレイアウト（構成）を示す。スクリーンディスプレイ１０００は、グラフィックス表示領域１００２とデータ表示領域１００４

30

＊【０１２７】ここで、遺伝子の発現状態は、存在（発現）、どちらともいえない、不存在（発現しない）のいずれかで示される。式「A and (X or Y) and (Q or R)」が成り立つ場合には、遺伝子が発現していると見なされる。言いかえると、Ｐ１＞＝２．１、Ｐ２＞＝０．２０、Ｐ３＞＝１．１が成り立つ場合には、遺伝子が発現していると考えられる。また、式「B and X and Q」が成り立つ場合にも、遺伝子が発現していると見なされる。

【０１２８】以上の説明に基づいて、遺伝子の発現状態を以下にまとめる。

＊【０１２９】

【０１３０】

【０１３１】

の２つの部分に分割される。グラフィックス表示領域には、ユーザーがデータを解釈する助けとなるグラフが表示される。また、データ表示領域には、ユーザーが遺伝子発現に関する基礎データを評価できるように、基礎データが表示される。

【０１３７】以下のスクリーンディスプレイの例に示すように、データ表示領域を縦欄と横列で表示される表形式にすることが望ましい。各縦欄には、その欄に含まれるデータを示す見出しがつけられる。また、各横列には、ある遺伝子に関する一つの実験あるいは実験の組み合わせから得られたデータが表示される。本明細書において、「実験」は、データを生成する工程を意味する。例えば、ハイブリッド結合されたチップの一つのイメージファイルから、多くの遺伝子に関する多くの「実験」データが生成される。また、実験は、異なったチップからデータを生成するものでもよい。

【０１３８】図２２は、選択された遺伝子の解析結果を示すスクリーンディスプレイである。スクリーンディスプレイ１０３０のグラフィックス表示領域には、選択された遺伝子の各塩基位置における完全対合プローブと不対合プローブのハイブリッド形成強度が棒グラフの形で示されている。選択された遺伝子は、データ表示領域１０３４に反転表示されている。

【０１３９】データ表示領域には、複数の縦欄の見出しが表示されている。「Experiment Name」は、その実験に関してユーザーが定義した名称を意味する。「Gene Name」は、その遺伝子の名称を意味する。「Positive」及び「Negative」の数は、図２０で説明したＮＰＯＳ及びＮＮＥＧの値を示す。「Pairs」は、その遺伝子の解析に用いられた完全符号プローブと不対合プローブのプ

ローブ対の数を示す。また、「Pos Fraction」は、遺伝子発現がポジティブであると判定されたプローブ対の割合（即ち、Positive／Pairs ）を示す。

【０１４０】「Avg Ratio 」欄は、ある遺伝子の全てのプローブに関するｉpm／Ｉmmの平均を示す。「Log A ％ 」欄は、ｌｏｇ（Ｉpm／ｉmm）の平均を示す。「PM Excess」欄は、ユーザーが規定した閾値より大きなハイブリッド形成強度を示す完全対合プローブの数を示す。「MM Excess 」欄は、ユーザーが規定した閾値より大きなハイブリッド形成強度を示す不対合プローブの数を示す。図２３に示される「Pos/Neg 」欄は、「Positive」欄と「Negative」欄の比を示す（「Negative」欄がゼロを含む場合には、「Inf」と表示される）。「Avg Diff」欄は、その遺伝子に関する平均強度差を示す。平均強度差は、図２０のステップ９７５で計算された値である（即ち、（ＩＤｉＦ）平均）。

【０１４１】「Abs Call」欄は、その実験に関する遺伝子発現コールを示す。この欄の値Ｐ．Ｍ、Ａは、それぞれ、存在、どちらともいえない、不存在を表す。好適な実施例の遺伝子発現コールに関する詳細は、図２０のステップ９７４を参照して既に説明した。

【０１４２】ユーザーが実験を選択すると、グラフィックス表示領域にユーザーがデータを解釈する助けとなるグラフが表示される。ユーザーは、ボタンバー１０３４を用いて、グラフィックス表示領域に表示されるべき一つあるいは複数のグラフを選択できる。さらに、所定の欄における値に従って、データ表示領域でデータを並べ変えることができる。

【０１４３】図２４は、選択された遺伝子の解析結果を表す別のスクリーンディスプレイを示す。スクリーンディスプレイ１０６０のグラフィックス表示領域１０６２には、各塩基位置における完全対合プローブと不対合プローブのハイブリッド形成強度の比のグラフが表示されている。ｘ軸は塩基位置を、ｙ軸は、ハイブリッド形成強度の比を表す。また、グラフには統計学的な比の閾値がプロットされているが、この例では、閾値は１．２である。ユーザーは、このグラフを用いて、閾値より上及び下のプローブ対の数（Ｉpm／ｉmm）を解析することができる。このグラフには、さらに遺伝子名と実験名が表示されている。

【０１４４】図２５は、選択された複数の遺伝子に関する実験結果の比較を表すスクリーンディスプレイを示す。スクリーンディスプレイ１１６０には、グラフィックス表示領域１０６２とデータ表示領域１１６４が含まれる。グラフィックス表示領域には、データ表示領域で選択された実験／遺伝子の各々に関して、各塩基位置における完全対合プローブと不対合プローブのハイブリッド形成強度の比のグラフが表示されている。好適な実施例では、選択された複数の遺伝子間の差異がはっきりとわかるように、実験名、遺伝子名及びデータプロットを遺伝子ごとに異なった色で示す。

【０１４５】図２６は、選択された複数の遺伝子に関する実験結果の比較を表す別のスクリーンディスプレイを示す。スクリーンディスプレイ１２００のグラフィックス表示領域１２０２には、データ表示領域１２０４で選択された遺伝子の発現レベルが表示されている。選択された遺伝子の発現レベルは、棒グラフの形で示されている。好適な実施例では、発現レベルは、平均強度差として表される（図２０の（ＩＤｉＦ）平均を参照）。このグラフには、さらに遺伝子名と実験名が表示されている。

【０１４６】図２７は、選択された複数の遺伝子に関する実験結果の比較を、グラフィックス表示領域に、いくつかのグラフの形で表示する、さらに別のスクリーンディスプレイを示す。スクリーンディスプレイ１２３０のグラフィックス表示領域１２３２には、データ表示領域１２３４で選択された遺伝子を解析した結果のグラフがいくつか示されている。この例では、選択された遺伝子に関する発現レベルグラフ１２３６、平均強度差グラフ１２３８、及び、ハイブリッド形成強度グラフ１２４０が表示されている。

【０１４７】図２８及び図２９は、基準走査データと実験走査データとを比較することにより、遺伝子の発現状態を判定する工程を示すフローチャートである。例えば、遺伝子が発現していることがわかっている生物検体から基準走査データをとるようにしてもよい。即ち、この走査データを別の生物検体と比較することにより、その遺伝子が発現しているかどうかを判定する。また、有機生命体において遺伝子の発現が時間とともにどのように変化するかを判定することもできる。ここで、「基準（基線）」は、標準点として用いられるものであることを意味する。

【０１４８】ステップ１３０２で、コンピュータシステムに、基準から得られたＮ対の完全対合プローブと不対合プローブの生の走査データが入力される。基準から得られた完全対合プローブのハイブリッド形成強度をｉpmで表し、同じく基準から得られた不対合プローブのハイブリッド形成強度をｉmmで表す。ステップ１３０４で、バックグラウンドの信号強度を各対の基準走査データのハイブリッド形成強度から差し引く。

【０１４９】一方、ステップ１３０６で、コンピュータシステムに、実験の標的である生物検体から得られたＮ対の完全対合プローブと不対合プローブの生の走査データが入力される。実験検体から得られた完全対合プローブのハイブリッド形成強度をＪpmで表し、同じく実験検体から得られた不対合プローブのハイブリッド形成強度をＪmmで表す。ステップ１３０８で、バックグラウンドの信号強度を各対の実験走査データのハイブリッド形成強度から差し引く。

【０１５０】次に、ステップ１３１０で、Ｉ－Ｊ対のハ

イブリッド形成強度を正規化する。例えば、ｉ－Ｊ対のハイブリッド形成強度を、対照プローブのハイブリッド形成強度で割るようにしてもよい。

【０１５１】ステップ１３１２で、プローブのＩ－Ｊ対のハイブリッド形成強度を差分閾値（ＤＤｉＦ）および比の閾値（ＲＤＩＦ）と比較する。具体的には、ある対（Ｊｐｍ－Ｊｍｍ）のハイブリッド形成強度と別の対（ｉｐｍ－ｉｍｍ）のハイブリッド形成強度の差が差分閾値以上であるかどうか、また、ある対（Ｊｐｍ－Ｊｍｍ）のハイブリッド形成強度と別の対（Ｉｐｍ－ｉｍｍ）のハイブリッド形成強度の比が比の閾値以上であるかどうかの判定がなされる。これらの閾値は、通常、一つあるいは複数の遺伝子の発現監視を精度よく行えるようにユーザーが規定した値である。

【０１５２】（Ｊｐｍ－Ｊｍｍ）－（ｉｐｍ－ｉｍｍ）＞＝ＤＤｉＦ、且つ、（Ｊｐｍ－Ｊｍｍ）／（ｉｐｍ－ｉｍｍ）＞＝ＲＤｉＦが成立する場合には、ステップ１３１４でＮＩＮＣを１増加させる。一般に、ＮＩＮＣは、実験検体のプローブ対の遺伝子発現が基準検体の遺伝子発現よりも大きい（即ち、増加している）可能性が高いことを示す値である。ＮＩＮＣを用いて、基準検体と比較して実験検体の遺伝子発現が大きい（即ち、増加している）か、小さい（即ち、減少している）か、あるいは、変化なしかの判定がなされる。

【０１５３】ステップ１３１６で、（Ｊｐｍ－Ｊｍｍ）－（Ｉｐｍ－ｉｍｍ）＞＝ＤＤＩＦ、且つ、（Ｊｐｍ－Ｊｍｍ）／（Ｉｐｍ－ｉｍｍ）＞＝ＲＤＩＦが成立するかどうかの判定がなされる。この式が成り立つ場合には、ＮＤＥＣを１増加させる。一般に、ＮＤＥＣは、実験検体のプローブ対の遺伝子発現が基準検体の遺伝子発現よりも小さい（即ち、減少している）可能性が高いことを示す値である。ＮＤＥＣを用いて、基準検体と比較して実験検体の遺伝子発現が大きい（即ち、増加している）か、小さい（即ち、減少している）か、あるいは、変化なしかの判定がなされる。

【０１５４】実験検体の遺伝子発現が大きいか、あるいは、小さいかを示すハイブリッド形成強度を有する各対のプローブに関して、ＮＰＯＳ、ＮＮＥＧ、及びＬＲ値を計算する。これらの値の計算に関しては、図２０を参照して前述した。それぞれの値の後ろにＢあるいはＥをつけて、その値が基準検体のものか実験検体のものかを示す。ステップ１３２２でハイブリッド形成強度を有する別のプローブ対が存在すると判定された場合には、上述と同様の方法で処理を行う。

【０１５５】次に、処理は図２９に移って、ステップ１３２４で、基準検体と実験検体の両方に関して、絶対決＊定演算を実行する。絶対決定演算は、基準検体と実験検体の各々に関して、遺伝子が発現しているか、どちらともいえないか、あるいは、発現していないか、を示すものである。従って、好適な実施例では、各検体に関して図２０のステップ９７２と９７４を実行することにより、このステップの処理が実現される。この処理を行うことによって、各検体の遺伝子発現状態が示される。

【０１５６】次のステップ１３２６では、決定行列を用いて、２つの検体間の遺伝子発現の差を求める。この決定行列では、上記のように計算された値Ｎ、ＮＰＯＳＢ、ＮＰＯＳＥ、ＮＮＥＧＢ、ＮＮＥＧＥ、ＮＩＮＣ、ＮＤＥＣ、ＬＲＢ、ＬＲＥが用いられる。決定行列の計算は、ＮＩＮＣがＮＤＥＣ以上であるかどうかによって変わる。即ち、以下のような計算が行われる。

【０１５７】ＮＩＮＣ＞＝ＮＤＥＣの場合には、次の４つのＰ値が求められる。

【０１５８】Ｐ１＝ＮＩＮＣ／ＮＤＥＣ
Ｐ２＝ＮＩＮＣ／Ｎ
Ｐ３＝（（ＮＰＯＳＥ－ＮＰＯＳＢ）－（ＮＮＥＧＥ－ＮＮＥＧＢ））／Ｎ
Ｐ４＝１０＊ＳＵＭ（ＬＲＥ－ＬＲＢ）／Ｎ

【０１５９】これらのＰ値を用いて、２つの検体間の遺伝子発現の差を求める。

【０１６０】上記と同様に、説明のために、Ｐ値を所定の範囲に分割する。Ｐ値は、全て、以下のような範囲に分割される。

【０１６１】Ａ＝（Ｐ１＞＝２．８）
Ｂ＝（２．８＞Ｐ１＞＝２．０）
Ｃ＝（Ｐ１＜２．０）

【０１６２】Ｘ＝（Ｐ２＞＝０．３４）
Ｙ＝（０．３４＞Ｐ２＞＝０．２４）
Ｚ＝（Ｐ２＜０．２４）

【０１６３】Ｍ＝（Ｐ３＞＝０．２０）
Ｎ＝（０．２０＞Ｐ３＞＝０．１２）
Ｏ＝（Ｐ３＜０．１２）

【０１６４】Ｑ＝（Ｐ４＞＝０．９）
Ｒ＝（０．９＞Ｐ４＞＝０．５）
Ｓ＝（Ｐ４＜０．５）

【０１６５】上記のプール値に従ってＰ値を所定の範囲に分割した後、２つの検体間の遺伝子発現の差を求める。

【０１６６】この場合、ＮＩＮＣ＞＝ＮＤＥＣであるため、遺伝子発現の変化は、増加、増加傾向、あるいは、変化なしで表される。以下に、遺伝子発現状態をまとめて示す。

【０１６７】

増加	A and (X or Y) and (Q or R) and (M or N or O)
	A and (X or Y) and (Q or R or S) and (M or N)
	B and (X or Y) and (Q or R) and (M or N)
	A and X and (Q or R or S) and (M or N or O)

【０１６８】

増加傾向　A or Y or S or O
B and (X or Y) and (Q or R) and O
B and (X or Y) and S and (M or N)
C and (X or Y) and (Q or R) and (M or N)

【０１６９】

変化なし　それ以外の全ての場合（例えば、あらゆる２の組み合わせ）

【０１７０】ユーザーへの出力は、増加をＩ、増加傾向をＭＩ、変化なしをＮＣのように示してもよい。

【０１７１】ＮＩＮＣ＜ＮＤＥＣの場合には、次の４つのＰ値が求められる。

【０１７２】Ｐ１＝ＮＤＥＣ／ＮＩＮＣ
Ｐ２＝ＮＤＥＣ／Ｎ
Ｐ３＝（（ＮＮＥＧＥ－ＮＮＥＧＢ）－（ＮＰＯＳＥ－ＮＰＯＳＢ））／Ｎ
Ｐ４＝１０＊ＳＵＭ（ＬＲＥ－ＬＲＢ）／Ｎ

【０１７３】これらのＰ値を用いて、２つの検体間の遺伝子発現の差を求める。

＊【０１７４】これらのＰ値は、上記のＮＩＮＣ＞＝ＮＤＥＣの場合と同じいくつかの範囲に分割する。Ｐ値は上述と同じ範囲に分割されるので、ここでは簡略化のために説明を繰り返さない。但し、以下に説明するように、これらの範囲は、通常、２つの検体間の遺伝子発現の異なった変化状態を示す。

【０１７５】この場合、ＮＩＮＣ＜ＮＤＥＣであるため、遺伝子発現の変化は、減少、減少傾向、あるいは、変化なしで表される。以下に、遺伝子発現状態をまとめて示す。

＊【０１７６】

減少　A and (X or Y) and (Q or R) and (M or N or O)
A and (X or Y) and (Q or R or S) and (M or N)
B and (X or Y) and (Q or R) and (M or N)
A and X and (Q or R or S) and (M or N or O)

【０１７７】

減少傾向　A or Y or S or O
B and (X or Y) and (Q or R) and O
B and (X or Y) and S and (M or N)
C and (X or Y) and (Q or R) and (M or N)

【０１７８】

変化なし　それ以外の全ての場合（例えば、あらゆる２の組み合わせ）

【０１７９】ユーザーへの出力は、減少をＤ、減少傾向をＤＩ、変化なしをＮＣのように示してもよい。

【０１８０】このようにして、基準検体と実験検体との間の遺伝子発現の相対的な差を求めることができる。（例えば、ステップ１３２４で）両方の検体において遺伝子が発現状態であると表示され、且つ、以下の式が全て成立する場合には、追加試験を行って、Ｉ、ＭＩ、ＤＭ、Ｄ（即ちＮＣ以外の）コールをＮＣに変更するように構成してもよい。

【０１８１】（ＩＤＩＦＢ）の平均＞＝２００
（ＩＤＩＦＥ）の平均＞＝２００
１．４＝｛（ＩＤＩＦＥ）の平均｝／｛（ＩＤＩＦＢ）の平均｝＞＝０．７

【０１８２】即ち、両方の検体で遺伝子が発現して場合に、各検体の平均強度差が比較的大きいか、あるいは、両方の検体の平均強度差がほとんど同じであれば、増加あるいは減少のコールは（それが傾向に留まっているかどうかに係わらず）変化なしコールに変更される。ＩＤＩＦＢ及びＩＤＩＦＥは、各検体に関するＩＤＩＦの絶和をＮで割ることにより計算される。

【０１８３】次に、ステップ１３２８で、定量的な差の評価を行うための値を計算する。各対に関して、（（Ｊpm－Ｊmm）－（Ｉpm－Ｉmm））の平均を計算する。また、Ｊpm－Ｊmmの平均とＩpm－Ｉmmの平均の比を計算する。ステップ１３３０で、これらの値を用いて、他の実験結果との比較を行うことができる。

【０１８４】図３０は、２つの実験における遺伝子発現の変化を表すスクリーンディスプレイを示す。スクリーンディスプレイ１４００には、グラフィックス表示領域１４０２とデータ表示領域１４０４が含まれる。ユーザーがある所定の遺伝子に関して２つの実験を選択することにより、その実験結果の比較が行われる。ここでは、説明をわかりやすくするために、基準データと実験データの２つを呼び出して、それらのデータの比較を行う場合を考える。例えば、ユーザーは、「ｇ１８２５０６」と名付けられた遺伝子に関して２つの実験を選択する。２つの実験結果の比較自体が実験であるため、ユーザーは、図３０のデータ表示領域に「foo」と表示されているように、実験名を入力することができる。図３１は、２つの実験における遺伝子発現の変化を表す別のスクリーンディスプレイを示す。

【０１８５】システムは、図３２で説明した方法に従っ

て．選択された２つの実験における遺伝子発現の変化を求める。データ表示領域には、この比較によって生成されたデータが表形式で示されている。「Experiment Name」は、比較実験に対してユーザーが規定した名称を示す。「Gene Name」は、遺伝子の名称を示す。「Inc」及び「Dec」の数値は、図２８を参照して説明したように．ＮｉＮＣおよびＮＤＥＣの値を示す。具体的には、「Inc」は、完全対合プローブと不対合プローブのハイブリッド形成強度の差及び比が実験データにおいて有意に大きいような遺伝子における塩基位置の数を示す。

【０１８６】「Inc Ratio」欄は、ハイブリッド形成強度が増加した塩基位置の数を解析対象の遺伝子における全ての塩基位置の数で割った値を示す。「Dec Ratio」欄は．ハイブリッド形成強度が減少した塩基位置の数を解析対象の遺伝子における全ての塩基位置の数で割った値を示す。「Pos Change」欄は、実験データと基準データにおいて正の得点を有するプローブ対の数の差を示す。「Neg Change」欄は．実験データと基準データにおいて負の得点を有するプローブ対（完全対合プローブと不対合プローブ）の数の差を示す。

【０１８７】「Inc/Dec」欄は、実験データにおいてハイブリッド形成強度が増加したプローブ対の数とハイブリッド形成強度が減少したプローブ対の数の比を示す。「Avg Diff」欄は、実験データにおける平均強度差を示す。

【０１８８】（図示されていない）「Diff Call」欄は．所定の遺伝子に関する２つの実験における発現レベルの変化を示す。この欄において、Ｉは遺伝子発現の増加を．ＭＩは遺伝子発現の増加傾向を．Ｄは遺伝子発現の減少を、ＭＤは遺伝子発現の減少傾向を、ＮＣは変化なしを、？は未知であることを、それぞれ示す。好適な実施例において．発現レベルの変化は．図２９のステップ１３２６を参照して説明したように計算される。

【０１８９】遺伝子発現の変化の演算に加えて．ユーザーは．データ解析のためにグラフを選択することもできる。グラフィックス表示領域１４０２には、基準データと実験データを示す３つのグラフが表示されている。

【０１９０】図３２は、２つの実験における遺伝子発現の変化を３次元棒グラフの形で示したスクリーンディスプレイを示す。スクリーンディスプレイ１４４０のグラフィックス表示領域１４４２には、データ表示領域１４４４で選択された遺伝子の発現レベルを示す３次元棒グラフが表示されている。ユーザーは．一つあるいは複数の遺伝子をデータ表示領域で選択して．これら遺伝子の発現レベルを示す３次元棒グラフを作成するようにシステムに命令する。好適な実施例において、発現レベルは．平均強度差（即ち、（ＩＤｉＦ）の平均）として与えられる。３次元棒グラフにより、ユーザーは．簡単に．複数の遺伝子の発現レベルを調べることができる。また．複数の実験結果から選択された．同じ遺伝子に関するデータを同時に示して回転させることにより．発現レベルの差を表示するようにすることもできる。

【０１９１】結語

上述の説明は例示を目的としたもので．何ら発明を限定するものではない。当業者に明らかなように、この開示に基づき、本発明を様々に変更することができる。例えば．本発明では．（天然のものも人工のものも含めて）ＤＮＡの評価に関して説明しているが．本発明の方法を．ＲＮＡのような他の物質が合成されたチップの解析に適用することも可能である。従って．本発明の範囲は．先の説明によって決まるものではなく、特許請求の範囲およびそれと等価と考えられる全範囲によって決定される。

【図面の簡単な説明】

【図１】本発明のソフトウェアを実行するために用いられるコンピュータシステムの例を示す図。

【図２】典型的なコンピュータシステムを示すシステムブロック図。

【図３】ＤＮＡやＲＮＡ等の生体物質の配列（アレイ）を形成し、解析するための全体的なシステムをを示す図。

【図４】全体的なシステムに関するソフトウェアの実施例を示す図。

【図５】全体的なシステムで形成されたチップの全体的な構成を示す図。

【図６】チップ上の核酸プローブが標識された標的に結合する様子を概念的に示す図。

【図７】チップ上に列で配置された核酸プローブを示す図。

【図８】図７のような参照配列と共に．チップ上の標的のハイブリッド形成パターンを示す図。

【図９】標準構造及び代替構造を示す図。

【図１０】チップから得られたハイブリッド形成強度を表すスクリーンディスプレイを示す図。

【図１１】関連するプローブのハイブリッド形成強度から塩基コールを演算する方法を示すフローチャート。

【図１２】関連するプローブのハイブリッド形成強度から塩基コールを演算する別の方法を示すフローチャート。

【図１３】一つのユニットグループにおいて塩基コール処理を行う方法を示すフローチャート。

【図１４】複数のユニットグループにおいて塩基コール処理を行う方法を示すフローチャート。

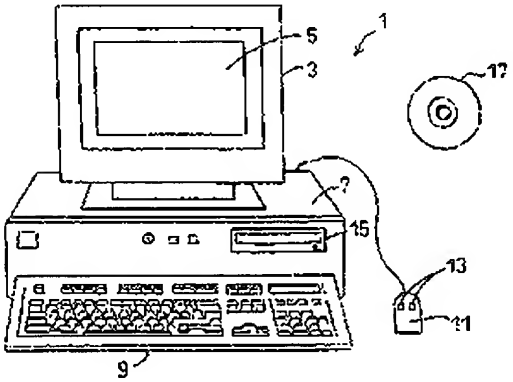
【図１５】一つのユニットグループに関して一つの塩基を呼び出す方法を示すフローチャート。

【図１６】塩基コール処理を実行するために最適のユニットグループを選択する方法を示すフローチャート。

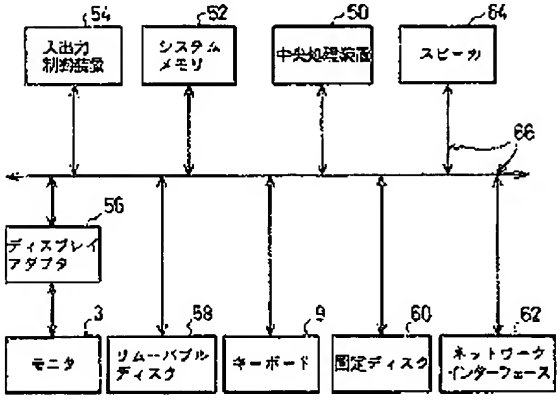
【図１７】一つあるいは複数のチップから得られた実験データに基づきヌクレオチドを解析するためのスクリーンディスプレイを示す図。

【図１８】一つあるいは複数のチップから得られた実験データに基づきヌクレオチドを解析するためのスクリーンディスプレイを示す図。
【図１９】完全対合プローブと不対合プローブ対のハイブリッド形成強度を比較することにより、ある遺伝子の発現状態を監視する方法を示すフローチャート。
【図２０】決定行列を用いて遺伝子が発現しているかどうかを判定する方法を示すフローチャート。
【図２１】遺伝子発現を監視するソフトウェアのスクリーンディスプレイの配置を示す図。
【図２２】選択された遺伝子の解析結果を表すスクリーンディスプレイを示す図。
【図２３】選択された遺伝子の解析結果を表すスクリーンディスプレイを示す図。
【図２４】選択された遺伝子の解析結果を表す別のスクリーンディスプレイを示す図。
【図２５】選択された複数の遺伝子に関する実験結果の比較を表すスクリーンディスプレイを示す図。
＊【図２６】選択された複数の遺伝子に関する実験結果の比較を表す別のスクリーンディスプレイを示す図。
【図２７】選択された複数の遺伝子に関する実験結果の比較をグラフィックス表示領域における複数のグラフを用いて表す別のスクリーンディスプレイを示す図。
【図２８】基準走査データと実験走査データとを比較することにより、ある遺伝子の発現状態を決定する方法を示すフローチャート。
【図２９】基準走査データと実験走査データとを比較することにより、ある遺伝子の発現状態を決定する方法を示すフローチャート。
【図３０】２つの実験における遺伝子発現の変化を表すスクリーンディスプレイを示す図。
【図３１】２つの実験における遺伝子発現の変化を表すスクリーンディスプレイを示す図。
【図３２】２つの実験における遺伝子発現の変化を３次元棒グラフの形で表すスクリーンディスプレイを示す図。

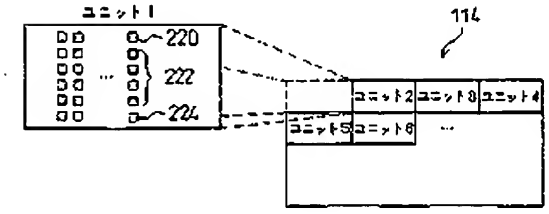
【図１】



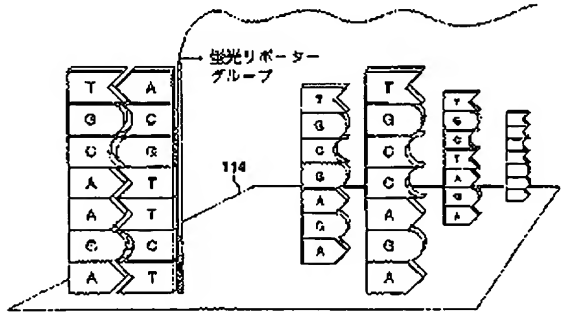
【図２】



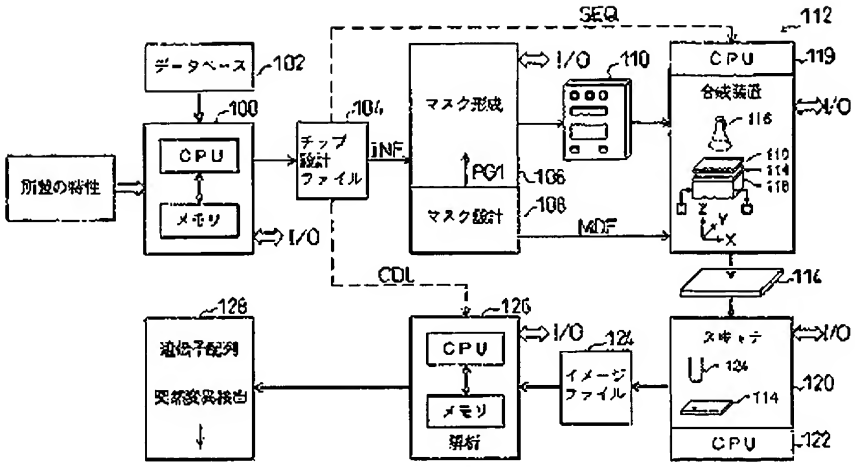
【図５】



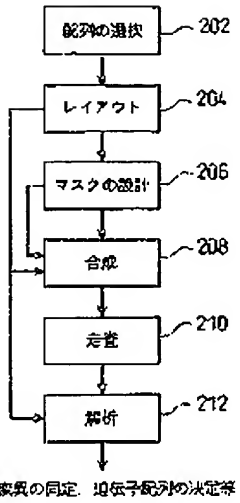
【図６】



【図３】

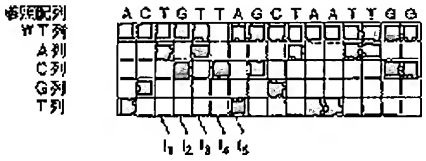


【図４】

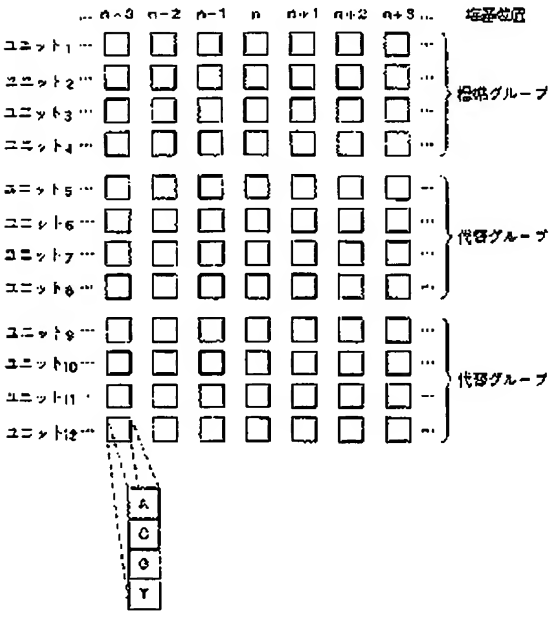


【図７】

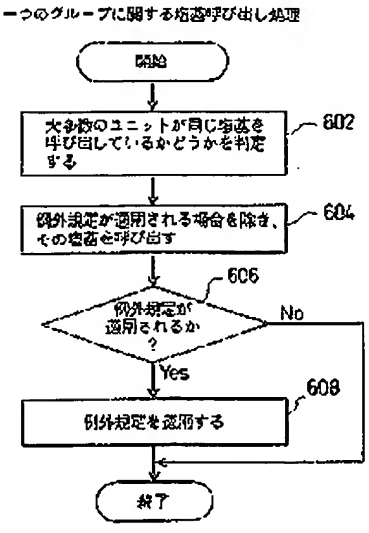
【図８】



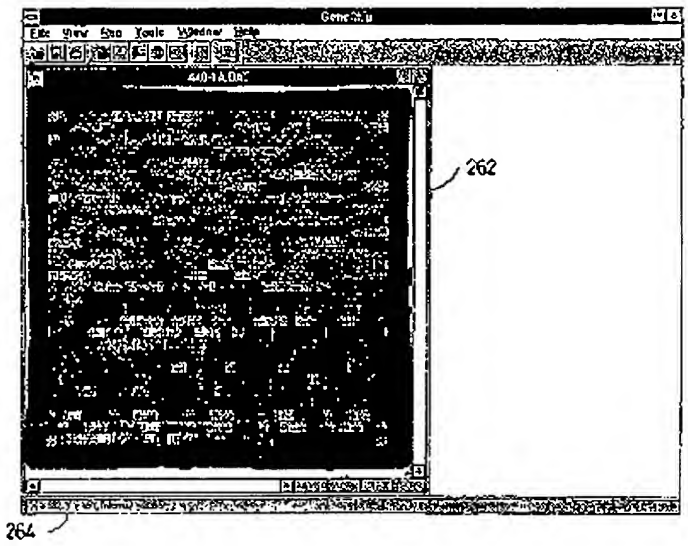
【図９】



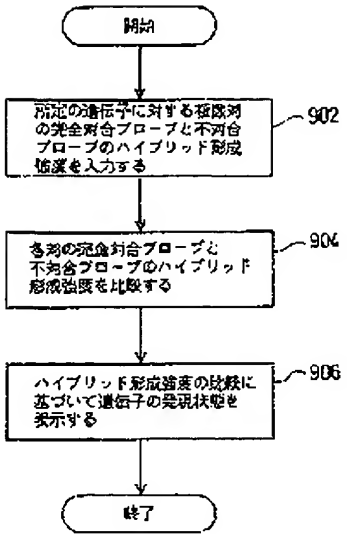
【図１５】



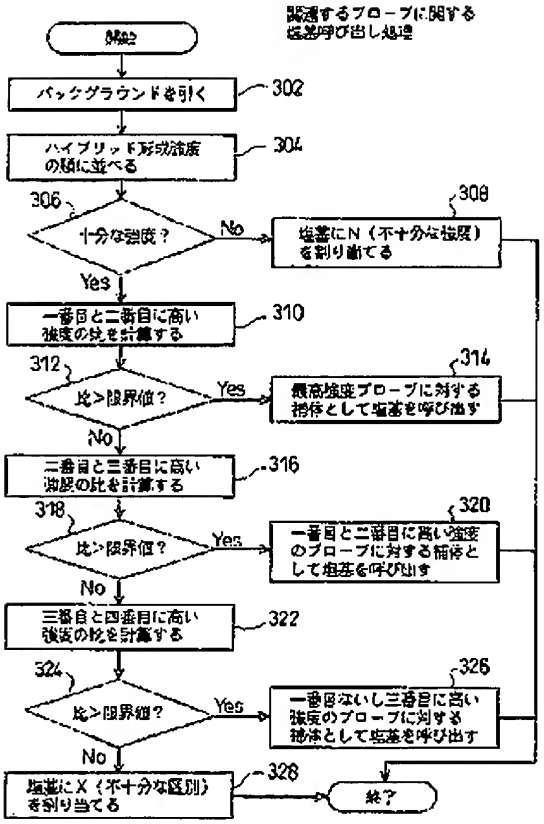
【図１０】



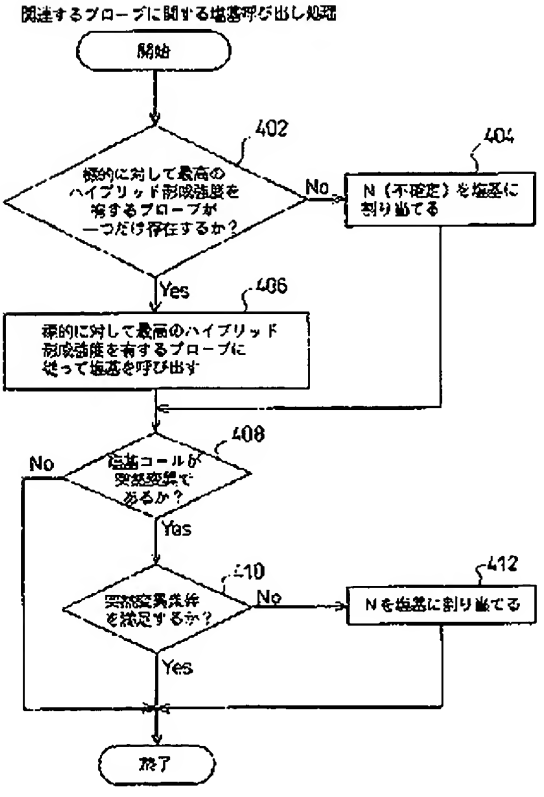
【図１９】



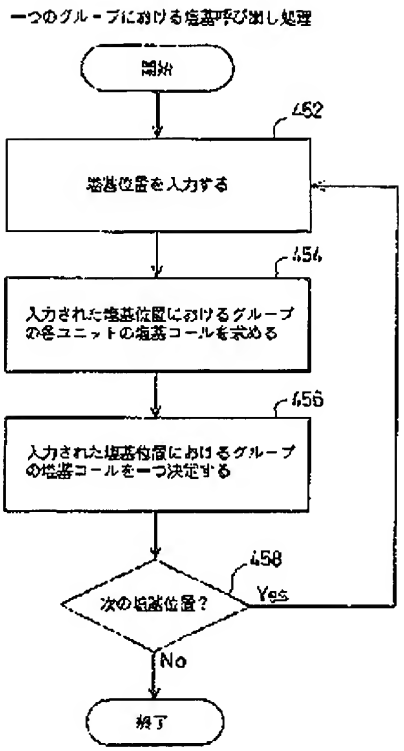
【図１１】



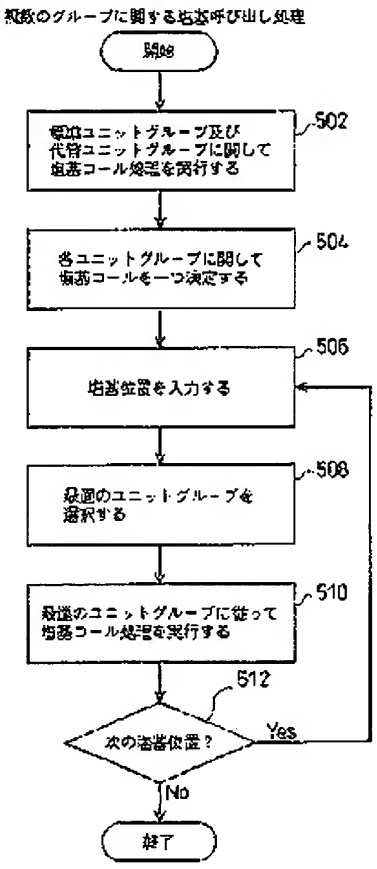
【図１２】



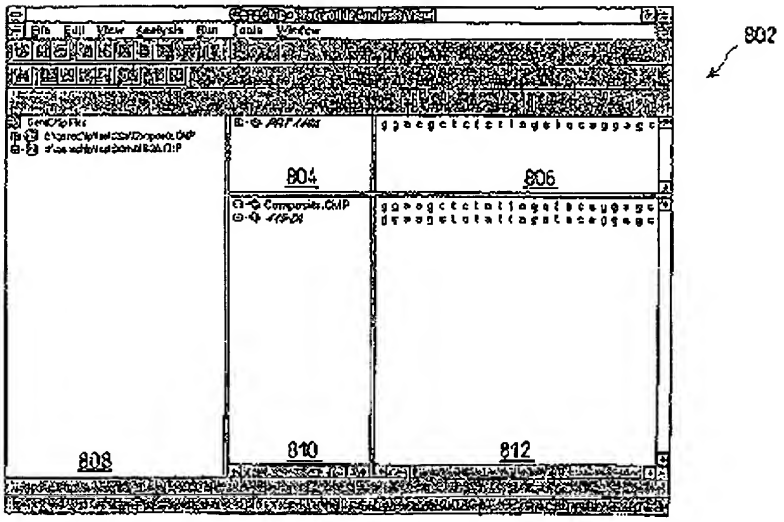
【図１３】


一つのグループにおける塔基呼び出し処理
開始
452
塔基位置を入力する
454
入力された塔基位置におけるグループの各ユニットの塔基コールを求める
456
入力された塔基位置におけるグループの塔基コールを一つ決定する
458
次の塔基位置？
Yes
No
終了

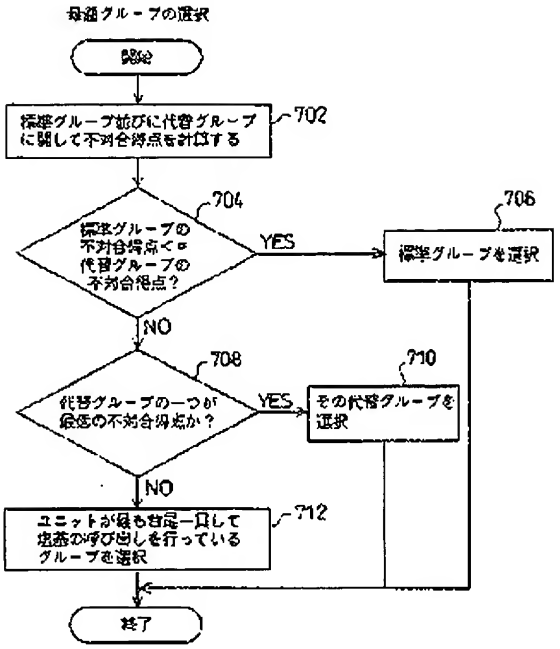
【図１４】


複数のグループに関する塔基呼び出し処理
開始
502
要塔ユニットグループ及び代替ユニットグループに関して塔基コール処理を実行する
504
各ユニットグループに関して塔基コールを一つ決定する
506
塔基位置を入力する
508
最適のユニットグループを選択する
510
最適のユニットグループに従って塔基コール処理を実行する
512
次の塔基位置？
Yes
No
終了

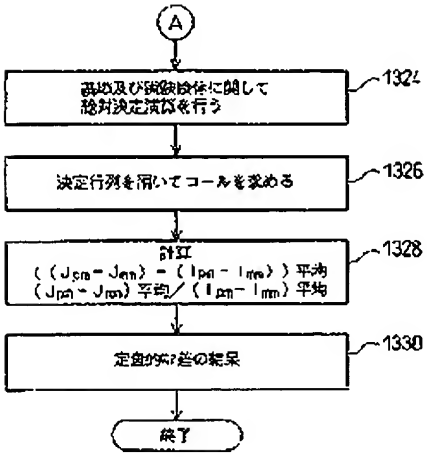
【図１７】


802
804
806
808
810
812

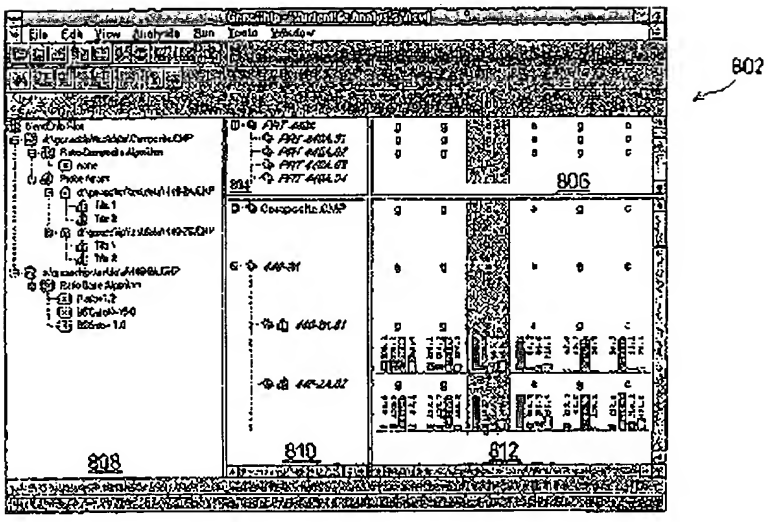
【図16】


最適グループの選択
開始
標準グループ並びに代替グループに関して不対合得点を計算する
702
704
標準グループの不対合得点＜＝代替グループの不対合得点？
YES
706
標準グループを選択
NO
708
代替グループの一つが最低の不対合得点か？
YES
710
その代替グループを選択
NO
ユニットが最も自足一貫して虚偽の呼び出しを行っているグループを選択
712
終了

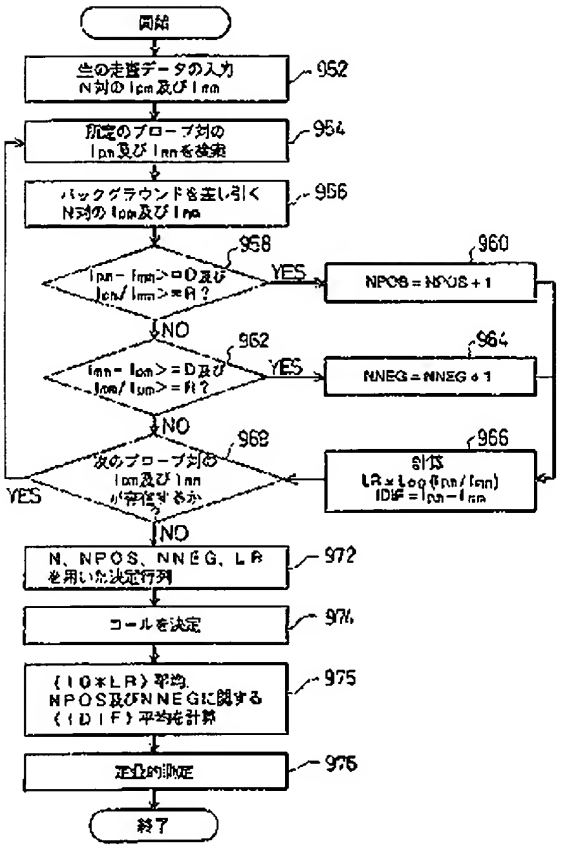
【図29】


A
1324
1326
決定行列を用いてコールを求める
1328
計算
1330
終了

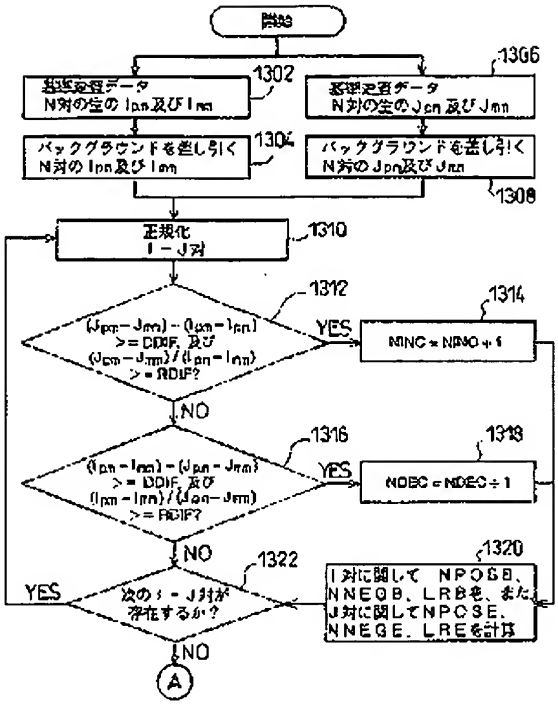
【図18】


802
806
808
810
812

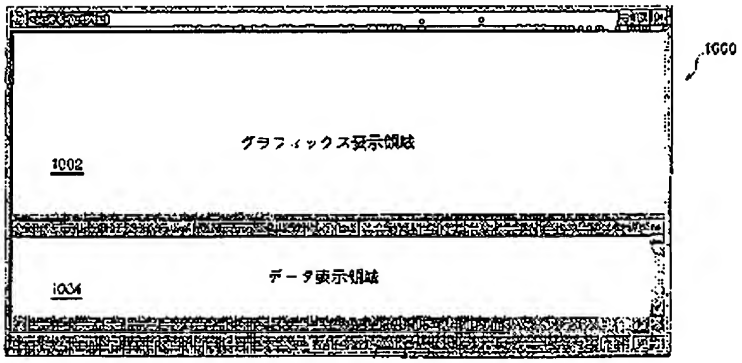
【図２０】


開始
当の走査データの入力
N対の Ipn及び Imn
952
所定のプローブ対の
Ipn及び Imnを検索
954
バックグラウンドを差し引く
N対の Ipn及び Imn
956
958
YES
NPOS = NPOS + 1
960
NO
962
YES
NNEG = NNEG + 1
964
NO
968
YES
次のプローブ対の
Ipn及び Imn
が存在するか
966
SIS
NO
N、NPOS、NNEG、LR
を用いた決定行列
972
コールを決定
974
（IQ＊LR）平均、
NPOS及びNNEGに関する
（IDIF）平均を計算
975
976
終了

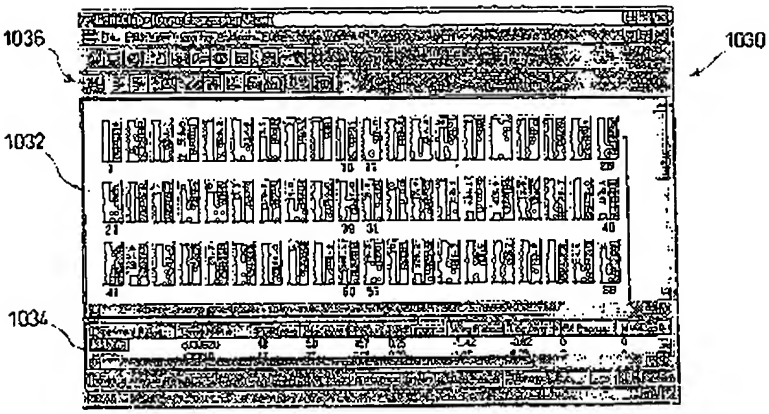
【図２８】


開始
1302
1306
N対の生の Ipn及び Imn
N対の生の Jpn及び Jmn
1304
バックグラウンドを差し引く
N対の Ipn及び Imn
バックグラウンドを差し引く
N対の Jpn及び Jmn
1308
正規化
I － J対
1310
1312
YES
1314
NINC = NINC + 1
NO
1316
YES
1318
NDEC = NDEC + 1
NO
1322
YES
1320
I対に関して NPOSB、
NNEGB、LRBを、また
J対に関してNPOSE、
NNEGE、LREを計算
NO
A

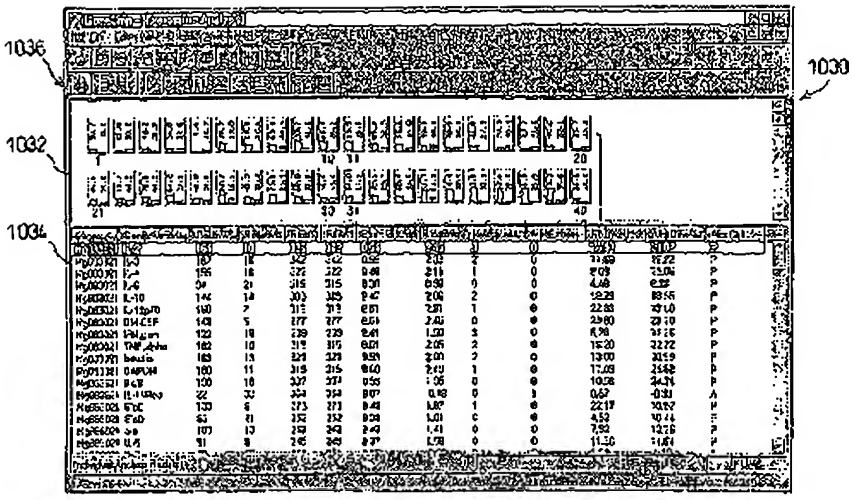
【図２１】


1000
グラフィックス表示領域
1002
データ表示領域
1004

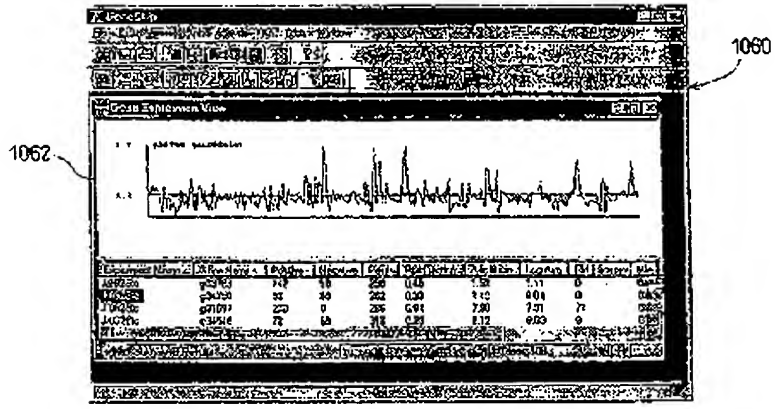
【図22】



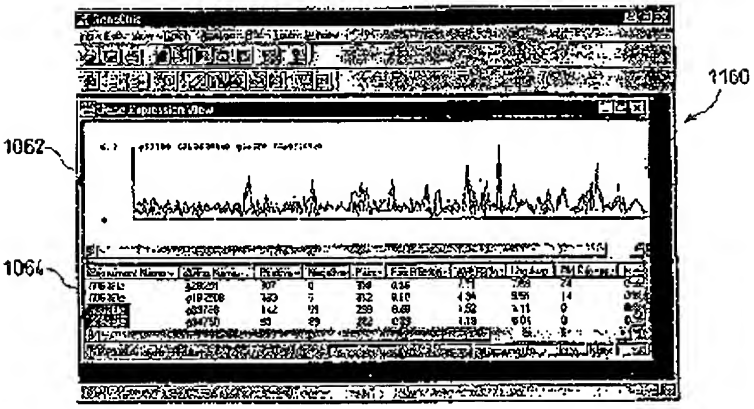
【図23】



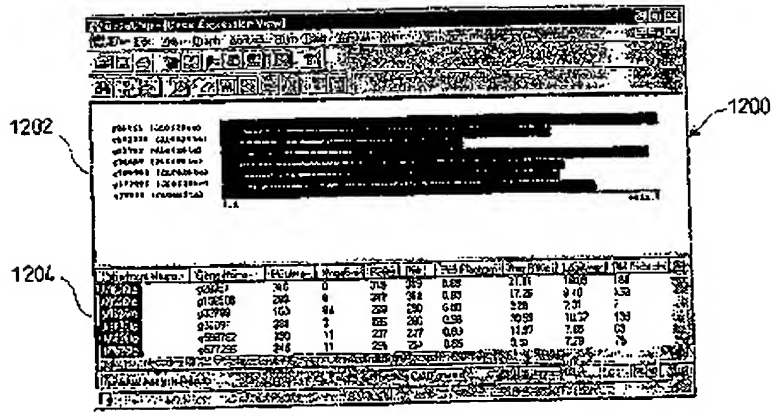
【図24】



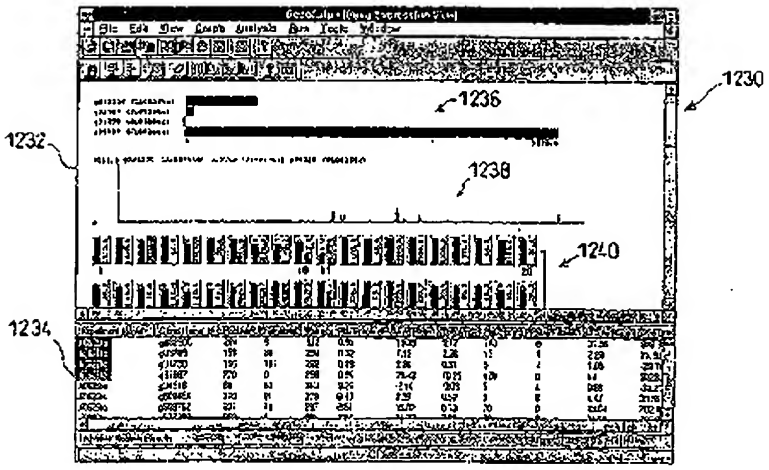
【図25】



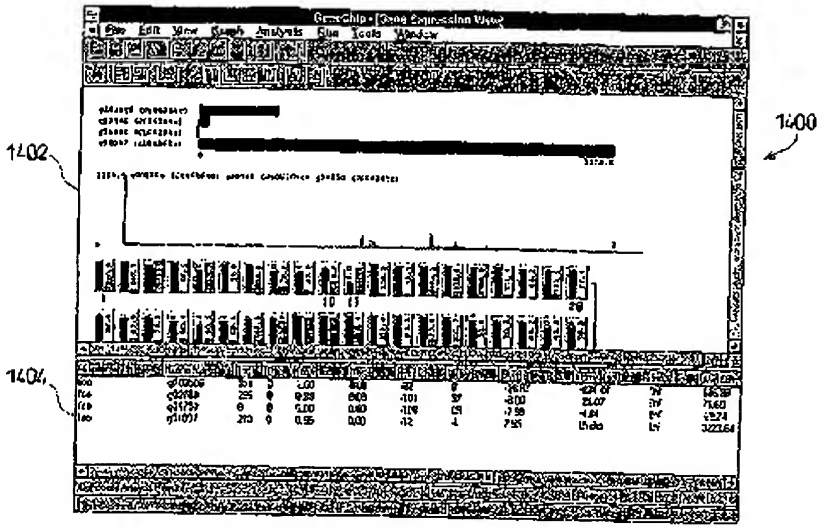
【図26】



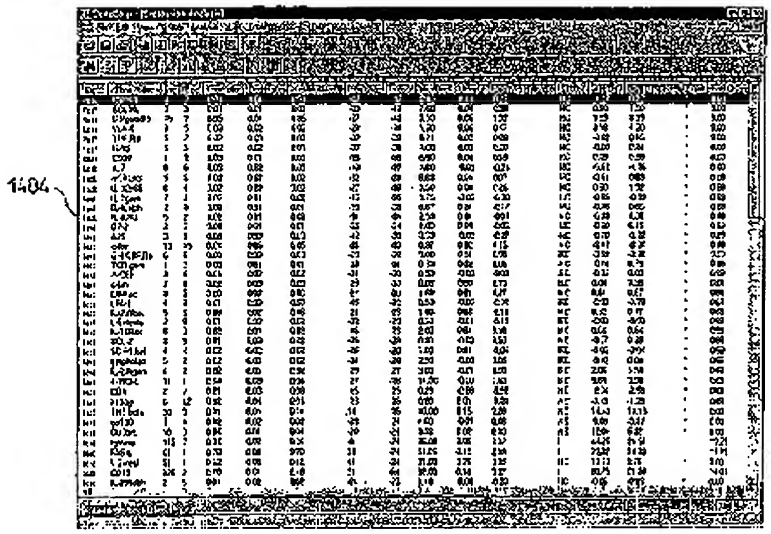
【図27】



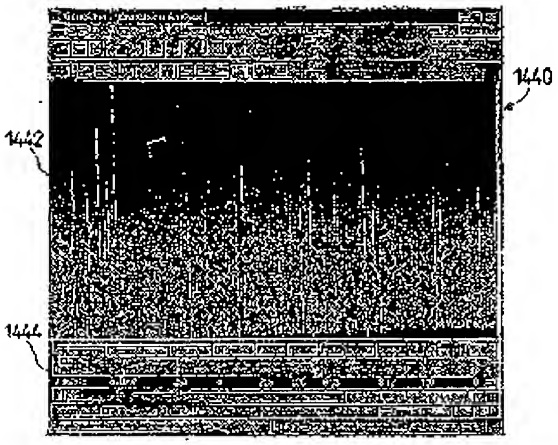
【図30】



【図31】



【図32】



フロントページの続き

(72)発明者　モリス・エス．・マクドナルド
アメリカ合衆国　カリフォルニア州94305　スタンフォード，ソノマ・テラス，843
(72)発明者　マイケル・ビー．・ミットマン
アメリカ合衆国　カリフォルニア州94303　パロ・アルト，セイント・フランシス・ドライブ，2377
(72)発明者　デビット・ジェイ．・ロックハート
アメリカ合衆国　カリフォルニア州94041　マウンテン・ビュー，マウンテン・ビュー・アベニュー，610
(72)発明者　ミン－スー・ホー
アメリカ合衆国　カリフォルニア州95133　サン・ホセ，フォックスリッジ・ウェイ，922

(72)発明者　デレク・バーンハート
アメリカ合衆国　カリフォルニア州94087
サニーヴェール，シラス・ウェイ，422

(72)発明者　ルイス・シー．・ジェボンス
アメリカ合衆国　カリフォルニア州94087
サニーヴェール，ロマーナ・アベニュー，　701

[外国語明細書]

1 Title of Invention

COMPUTER-AIDED TECHNIQUES FOR ANALYZING BIOLOGICAL SEQUENCES

2 Claims

1. In a computer system, a method of analyzing a sample nucleic acid sequence, the method comprising the steps of:

inputting a plurality of base calls for each base position along at least a portion of the sample nucleic acid sequence;

for each base position, analyzing the plurality of base calls to generate a single base call; and

displaying single base calls for base positions along the at least a portion of said sample nucleic acid sequence, each of the single base calls being derived from the plurality of base calls for a specific base position in the sample nucleic acid sequence.

2. The method of claim 1, wherein the analyzing step comprises the steps of:

for each base position, determining a base call of the plurality of base calls which occurs most often; and

generating the single base call as the base call that occurs most often at the base position.

3. The method of claim 1, further comprising the step of displaying a screen icon which when activated by a user causes the plurality of base calls at each base position to be displayed.

4. The method of claim 1, further comprising the step of displaying a screen icon which when activated by a user causes the plurality of base calls at each base position not to be displayed.

5. The method of claim 1, further comprising the step of displaying the plurality of base calls at each base aligned with the single base calls according to base position.

6. The method of claim 5, further comprising the step of displaying with each base call of the plurality of base calls hybridization intensities indicating hybridization affinity of a probe and the sample nucleic acid sequence,

wherein each base call is determined by an analysis of the hybridization intensities.

7. In a computer system, a method of calling an unknown base in a sample nucleic acid sequence, the method comprising the steps of:

receiving hybridization intensities for a plurality of sets of nucleic acid probes, each hybridization intensity indicating a hybridization affinity between a nucleic acid probe and the sample nucleic acid sequence;

computing a base call for the unknown base for each set of probes; and

computing a single base call for the plurality of sets of probes according to the base call for the unknown base which occurs most often for the plurality of sets of probes.

8. The method of claim 7, wherein each set of probes was generated according to a same reference sequence.

9. The method of claim 7, further comprising the step of checking exception rules that specify the single base call for the plurality of sets of nucleic acid probes under certain conditions.

10. In a computer system, a method of dynamically changing parameters for a computer-implemented base calling procedure, the method comprising the steps of:

generating base calls for at least a portion of a sample nucleic acid sequence utilizing the base calling procedure, the base calling procedure including a parameter that is changeable by a user;

displaying the base calls for the at least a portion of a sample nucleic acid sequence;

displaying the parameter of the base calling procedure;

receiving input from the user specifying a new value for the parameter of the base calling procedure;

generating updated base calls for the at least a portion of a sample nucleic acid sequence utilizing the base calling procedure and the new value for the parameter; and

displaying the updated base calls for the at least a portion of a sample nucleic acid sequence.

11. The method of claim 10, further comprising the step of displaying a plurality of user-changeable parameters for the base calling procedure.

12. The method of claim 10, wherein the parameter is selected from the group consisting of a constant, threshold, and range.

13. In a computer system, a method of monitoring expression of a gene in a sample nucleic acid sequence, the method comprising the steps of:

inputting a plurality of hybridization intensities of pairs of perfect match and mismatch probes, the perfect match probes being perfectly complementary to the gene and the mismatch probes having at least one base mismatch with the gene, and the hybridization intensities indicating hybridization infinity between the perfect match and mismatch probes and the sample nucleic acid sequence;

comparing the hybridization intensities of each pair of perfect match probes in order to generate a gene expression call of the sample nucleic acid sequence; and

displaying the gene expression call.

14. The method of claim 13, further comprising the step of comparing a difference between hybridization intensities of perfect match and mismatch probes at a base position to a difference threshold.

15. The method of claim 13, further comprising the step of comparing a quotient of hybridization intensities of perfect match and mismatch probes at a base position to a ratio threshold.

16. The method of claim 13, further comprising the step of utilizing a decision matrix to determine the gene expression call.

17. The method of claim 13, wherein the gene expression call is selected from the group consisting of expressed, marginal, and absent.

18. In a computer system, a method of monitoring expression of a gene in a sample nucleic acid sequence, the method comprising the steps of:

inputting a plurality of hybridization intensities of pairs of perfect match and mismatch probes, the perfect match probes being perfectly complementary to the gene and the mismatch probes having at least one base mismatch with the gene, and the hybridization intensities indicating hybridization infinity between the perfect match and mismatch probes and the sample nucleic acid sequence;

comparing the hybridization intensities of each pair of perfect match

probes; and

generating a gene expression call of the sample nucleic acid sequence.

19. The method of claim 18, further comprising the step of comparing a difference between hybridization intensities of perfect match and mismatch probes at a base position to a difference threshold.

20. The method of claim 18, further comprising the step of comparing a quotient of hybridization intensities of perfect match and mismatch probes at a base position to a ratio threshold.

21. The method of claim 18, further comprising the step of utilizing a decisinon matrix to determine the gene expression call.

22. The method of claim 18, wherein the gene expression call is selected from the group consisting of expressed, marginal, and absent.

23. In a computer system, a method of monitoring change in expression of a gene in a sample nucleic acid sequence, the method comprising the steps of:

inputting a plurality of hybridization intensities of pairs of perfect match and mismatch probes, the perfect match probes being perfectly complementary to the gene and the mismatch probes having at least one base mismatch with the gene, and the hybridization intensities indicating hybridization infinity between the perfect match and mismatch probes and the sample nucleic acid sequence;

comparing the hybridization intensities of each pair of perfect match probes in order to generate a gene expression level of the sample nucleic acid sequence;

determining a change in expression by comparing the gene expression level to a baseline gene expression level; and

displaying the change in expression of the gene in the sample nucleic acid.

24. The method of claim 23, wherein the change in expression is displayed as a graph.

25. The method of claim 23, further comprising the step of generating the baseline expression level according to the inputting and comparing

steps of claim 23.

26. The method of claim 23, further comprising the step of comparing hybridization intensities of perfect match and mismatch probes hybridizing with the sample nucleic acid sequence and hybridization intensities of perfect match and mismatch probes hybridizing with a baseline sequence to a difference threshold.

27. The method of claim 23, further comprising the step of comparing hybridization intensities of perfect match and mismatch probes hybridizing with the sample nucleic acid sequence and hybridization intensities of perfect match and mismatch probes hybridizing with a baseline sequence to a ratio threshold.

28. The method of claim 23, further comprising the step of utilizing a decision matrix to determine the change in expression of the gene in the sample nucleic acid.

29. The method of claim 23, wherein the change in expression of the gene in the sample nucleic acid is selected from the group consisting of increased, marginal increase, decreased, marginal decrease, and no change.

30. In a computer system, a method of monitoring change in expression of a gene in a sample nucleic acid sequence, the method comprising the steps of:

inputting a plurality of hybridization intensities of pairs of perfect match and mismatch probes, the perfect match probes being perfectly complementary to the gene and the mismatch probes having at least one base mismatch with the gene, and the hybridization intensities indicating hybridization infinity between the perfect match and mismatch probes and the sample nucleic acid sequence;

comparing the hybridization intensities of each pair of perfect match probes in order to generate a gene expression level of the sample nucleic acid sequence; and

determining a change in expression by comparing the gene expression level to a baseline gene expression level.

31. The method of claim 30, further comprising the step of generating the baseline expression level according to the inputting and comparing

steps of claim 30.

32. The method of claim 30, further comprising the step of comparing hybridization intensities of perfect match and mismatch probes hybridizing with the sample nucleic acid sequence and hybridization intensities of perfect match and mismatch probes hybridizing with a baseline sequence to a difference threshold.

33. The method of claim 30, further comprising the step of comparing hybridization intensities of perfect match and mismatch probes hybridizing with the sample nucleic acid sequence and hybridization intensities of perfect match and mismatch probes hybridizing with a baseline sequence to a ratio threshold.

34. The method of claim 30, further comprising the step of utilizing a decision matrix to determine the change in expression of the gene in the sample nucleic acid.

35. The method of claim 30, wherein the change in expression of the gene in the sample nucleic acid is selected from the group consisting of increased, marginal increase, decreased, marginal decrease, and no change.

3 Detailed Description of Invention

The present invention relates to the field of computer systems. More specifically, the present invention relates to computer systems for analyzing biological sequences such as nucleic acid sequences.

Devices and computer systems for forming and using arrays of materials on a substrate are known. For example, PCT application WO92/10588, incorporated herein by reference for all purposes, describes techniques for sequencing or sequence checking nucleic acids and other materials. Arrays for performing these operations may be formed in arrays according to the methods of, for example, the pioneering techniques disclosed in U.S. Patent No. 5,143,854 and U.S. Patent Application No. 08/249,188, both incorporated herein by reference for all purposes.

According to one aspect of the techniques described therein, an array of nucleic acid probes is fabricated at known locations on a substrate or chip. A

fluorescently labeled nucleic acid is then brought into contact with the chip and a scanner generates an image file (which is processed into a cell file) indicating the locations where the labeled nucleic acids bound to the chip. Based upon the cell file and identities of the probes at specific locations, it becomes possible to extract information such as the monomer sequence of DNA or RNA. Such systems have been used to form, for example, arrays of DNA that may be used to study and detect mutations relevant to cystic fibrosis, the P53 gene (relevant to certain cancers), HIV, and other genetic characteristics.

Innovative computer-aided techniques for base calling are disclosed in U.S. Patent Application Nos. 08/531,137 (attorney docket no. 16528X-008210), 08/528,656 (attorney docket no. 16528X-017600), and 08/618,834 (attorney docket no. 16528X-016400), which are all hereby incorporated by reference for all purposes. However, improved computer systems and methods are still needed to evaluate, analyze, and process the vast amount of information now used and made available by these pioneering technologies.

Additionally, there is a need for improved computer-aided techniques for monitoring gene expression. Many disease states are characterized by differences in the expression levels of various genes either through changes in the copy number of the genetic DNA or through changes in levels of transcription (*e.g.*, through control of initiation, provision of RNA precursors, RNA processing, etc.) of particular genes. For example, losses and gains of genetic material play an important role in malignant transformation and progression. Furthermore, changes in the expression (transciption) levels of particular genes (*e.g.*, oncogenes or tumor suppressors), serve as signposts for the presence and progression of various cancers.

Similarly, control of the cell cycle and cell development, as well as diseases, are characterized by the variations in the transcription levels of particular genes. Thus, for example, a viral infection is often characterized by the elevated expression of genes of the particular virus. For example, outbreaks of *Herpes simplex*, Epstein-Barr virus infections (*e.g.*, infectious mononucleosis), cytomegalovirus, Varicella-zoster virus infections, parvovirus infections, human papillomavirus infections, *etc.* are all characterized by elevated expression of

various genes present in the respective virus. Detection of elevated expression levels of characteristic viral genes provides an effective diagnostic of the disease state. In particular, viruses such as herpes simplex, enter quiescent states for periods of time only to erupt in brief periods of rapid replication. Detection of expression levels of characteristic viral genes allows detection of such active proliferative (and presumably infective) states.

SUMMARY OF THE INVENTION

The present invention provides innovative systems and methods for analyzing biological sequences such as nucleic acid sequences. The computer system may analyze hybridization intensities indicating hybridization affinity between nucleic acid probes and a sample nucleic acid sequence in order to call bases in the sample sequence. Multiple base calls may be combined to form a single base call. Additionally, the computer system may analyze hybridization intensities in order to monitor gene expression or the change in gene expression as compared to a baseline.

According to one aspect of the invention, a computer-implemented method of calling an unknown base in a sample nucleic acid sequence comprises the steps of: receiving hybridization intensities for a plurality of sets of nucleic acid probes, each hybridization intensity indicating a hybridization affinity between a nucleic acid probe and the sample nucleic acid sequence; computing a base call for the unknown base for each set of probes; and computing a single base call for the plurality of sets of probes according to the base call for the unknown base which occurs most often for the plurality of sets of probes. Typically, the single base call is displayed on a screen display and a user is afforded the opportunity to display or not display the base cases from which the single base call is derived.

According to another aspect of the invention, a method of dynamically changing parameters for a computer-implemented base calling procedure comprises the steps of: generating base calls for at least a portion of a sample nucleic acid sequence utilizing the base calling procedure, the base calling procedure including a parameter that is changeable by a user; displaying the base calls for the at least a portion of a sample nucleic acid sequence; displaying the parameter of the base

calling procedure; receiving input from the user specifying a new value for the parameter of the base calling procedure; generating updated base calls for the at least a portion of a sample nucleic acid sequence utilizing the base calling procedure and the new value for the parameter; and displaying the updated base calls for the at least a portion of a sample nucleic acid sequence. Typically the user-changeable parameter is a constant, threshold, or range.

According to another aspect of the invention, a computer-implemented method of monitoring expression of a gene in a sample nucleic acid sequence comprises the steps of: inputting a plurality of hybridization intensities of pairs of perfect match and mismatch probes, the perfect match probes being perfectly complementary to the gene and the mismatch probes having at least one base mismatch with the gene, and the hybridization intensities indicating hybridization infinity between the perfect match and mismatch probes and the sample nucleic acid sequence; comparing the hybridization intensities of each pair of perfect match probes; and generating a gene expression call of the sample nucleic acid sequence. In preferred embodiments, the expression call is denoted as expressed, marginal, or absent.

According to another aspect of the invention, a computer-implemented method of monitoring change in expression of a gene in a sample nucleic acid sequence comprises the steps of: inputting a plurality of hybridization intensities of pairs of perfect match and mismatch probes, the perfect match probes being perfectly complementary to the gene and the mismatch probes having at least one base mismatch with the gene, and the hybridization intensities indicating hybridization infinity between the perfect match and mismatch probes and the sample nucleic acid sequence; comparing the hybridization intensities of each pair of perfect match probes in order to generate a gene expression level of the sample nucleic acid sequence; and determining a change in expression by comparing the gene expression level to a baseline gene expression level. The change in expression may be displayed as a graph on the display screen.

A further understanding of the nature and advantages of the inventions herein may be realized by reference to the remaining portions of the specification and the attached drawings.

DESCRIPTION OF PREFERRED EMBODIMENTS

General

The present invention provides innovative methods of identifying nucleotides (*i.e.*, base calling) in sample nucleic acid sequences and monitoring gene expression. In the description that follows, the invention will be described in reference to preferred embodiments. However, the description is provided for purposes of illustration and not for limiting the spirit and scope of the invention.

Fig. 1 illustrates an example of a computer system that may be used to execute software embodiments of the present invention. Fig. 1 shows a computer system 1 which includes a monitor 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have one or more buttons such as mouse buttons 13. Cabinet 7 houses a CD-ROM drive 15 and a hard drive (not shown) that may be utilized to store and retrieve software programs including computer code incorporating the present invention. Although a CD-ROM 17 is shown as the computer readable medium, other computer readable media including floppy disks, DRAM, hard drives, flash memory, tape, and the like may be utilized. Cabinet 7 also houses familiar computer components (not shown) such as a processor, memory, and the like.

Fig. 2 shows a system block diagram of computer system 1 used to execute software embodiments of the present invention. As in Fig. 1, computer system 1 includes monitor 3 and keyboard 9. Computer system 1 further includes subsystems such as a central processor 50, system memory 52, I/O controller 54, display adapter 56, removable disk 58, fixed disk 60, network interface 62, and speaker 64. Removable disk 58 is representative of removable computer readable media like floppies, tape, CD-ROM, removable hard drive, flash memory, and the like. Fixed disk 60 is representative of an internal hard drive or the like. Other computer systems suitable for use with the present invention may include additional or fewer subsystems. For example, another computer system could include more than one processor 50 (*i.e.*, a multi-processor system) or memory cache.

Arrows such as 66 represent the system bus architecture of computer

system 1. However, these arrows are illustrative of any interconnection scheme serving to link the subsystems. For example, display adapter 56 may be connected to central processor 50 through a local bus or the system may include a memory cache. Computer system 1 shown in Fig. 2 is but an example of a computer system suitable for use with the present invention. Other configurations of subsystems suitable for use with the present invention will be readily apparent to one of ordinary skill in the art. In one embodiment, the computer system is a workstation from Sun Microsystems.

The VLSIPS™ technology provides methods of making very large arrays of oligonucleotide probes on very small chips. *See* U.S. Patent No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, each of which is hereby incorporated by reference for all purposes. The oligonucleotide probes on the chip are used to detect complementary nucleic acid sequences in a sample nucleic acid of interest (the "target" nucleic acid).

The present invention provides methods of analyzing hybridization intensity files for a chip containing hybridized nucleic acid probes. In a representative embodiment, the files represent fluorescence data from a biological array, but the files may also represent other data such as radioactive intensity data. Therefore, the present invention is not limited to analyzing fluorescent measurements of hybridizations but may be readily utilized to analyze other measurements of hybridization.

For purposes of illustration, the present invention is described as being part of a computer system that designs a chip mask, synthesizes the probes on the chip, labels the nucleic acids, and scans the hybridized nucleic acid probes. Such a system is fully described in U.S. Patent Application No. 08/249,188 which is hereby incorporated by reference for all purposes. However, the present invention may be used separately from the overall system for analyzing data generated by such systems, such as at remote locations.

Fig. 3 illustrates a computerized system for forming and analyzing arrays of biological materials such as RNA or DNA. A computer 100 is used to design arrays of biological polymers such as RNA or DNA. The computer 100 may be, for example, an appropriately programmed IBM personal computer compatible

running Windows NT including appropriate memory and a CPU as shown in Figs. 1 and 2. The computer system 100 obtains inputs from a user regarding characteristics of a gene of interest, and other inputs regarding the desired features of the array. Optionally, the computer system may obtain information regarding a specific genetic sequence of interest from an external or internal database 102 such as GenBank. The output of the computer system 100 is a set of chip design computer files 104 in the form of, for example, a switch matrix, as described in PCT application WO 92/10092, and other associated computer files.

The chip design files are provided to a system 106 that designs the lithographic masks used in the fabrication of arrays of molecules such as DNA. The system or process 106 may include the hardware necessary to manufacture masks 110 and also the necessary computer hardware and software 108 necessary to lay the mask patterns out on the mask in an efficient manner. As with the other features in Fig. 3, such equipment may or may not be located at the same physical site, but is shown together for ease of illustration in Fig. 3. The system 106 generates masks 110 or other synthesis patterns such as chrome-on-glass masks for use in the fabrication of polymer arrays.

The masks 110, as well as selected information relating to the design of the chips from system 100, are used in a synthesis system 112. Synthesis system 112 includes the necessary hardware and software used to fabricate arrays of polymers on a substrate or chip 114. For example, synthesizer 112 includes a light source 116 and a chemical flow cell 118 on which the substrate or chip 114 is placed. Mask 110 is placed between the light source and the substrate/chip, and the two are translated relative to each other at appropriate times for deprotection of selected regions of the chip. Selected chemical reagents are directed through flow cell 118 for coupling to deprotected regions, as well as for washing and other operations. All operations are preferably directed by an appropriately programmed computer 119, which may or may not be the same computer as the computer(s) used in mask design and mask making.

The substrates fabricated by synthesis system 112 are optionally diced into smaller chips and exposed to marked targets. The targets may or may not be complementary to one or more of the molecules on the substrate. The targets are

marked with a label such as a fluorescein label (indicated by an asterisk in Fig. 3) and placed in scanning system 120. Scanning system 120 again operates under the direction of an appropriately programmed digital computer 122, which also may or may not be the same computer as the computers used in synthesis, mask making, and mask design. The scanner 120 includes a detection device 124 such as a confocal microscope or CCD (charge-coupled device) that is used to detect the location where labeled target (*) has bound to the substrate. The output of scanner 120 is an image file(s) 124 indicating, in the case of fluorescein labeled target, the fluorescence intensity (photon counts or other related measurements, such as voltage) as a function of position on the substrate. Since higher photon counts will be observed where the labeled target has bound more strongly to the array of polymers, and since the monomer sequence of the polymers on the substrate is known as a function of position, it becomes possible to determine the sequence(s) of polymer(s) on the substrate that are complementary to the target.

The image file 124 is provided as input to an analysis system 126 that incorporates the visualization and analysis methods of the present invention. Again, the analysis system may be any one of a wide variety of computer system(s). The present invention provides various methods of analyzing the chip design files and the image files, providing appropriate output 128. The present invention may further be used to identify specific mutations in a target such as DNA or RNA.

Fig. 4 provides a simplified illustration of the overall software system used in the operation of one embodiment of the invention. As shown in Fig. 4, the system first identifies the genetic sequence(s) or targets that would be of interest in a particular analysis at step 202. The sequences of interest may, for example, be normal or mutant portions of a gene, genes that identify heredity, or provide forensic information. Sequence selection may be provided via manual input of text files or may be from external sources such as GenBank. At step 204 the system evaluates the gene to determine or assist the user in determining which probes would be desirable on the chip, and provides an appropriate "layout" on the chip for the probes.

The chip usually includes probes that are complementary to a reference nucleic acid sequence which has a known sequence. A wild-type probe

is a probe that will ideally hybridize with the reference sequence and thus a wild-type gene (also called the chip wild-type) would ideally hybridize with wild-type probes on the chip. The target sequence is substantially similar to the reference sequence except for the presence of mutations, insertions, deletions, and the like. The layout implements desired characteristics such as arrangement on the chip that permits "reading" of genetic sequence and/or minimization of edge effects, ease of synthesis, and the like.

Fig. 5 illustrates the global layout of a chip. Chip 114 is composed of multiple units where each unit may contain different tilings for the wild-type sequence or multiple wild-type sequences. Unit 1 is shown in greater detail and shows that each unit is composed of multiple cells which are areas on the chip that may contain probes. Conceptually, each unit includes multiple sets of related cells. As used herein, the term cell refers to a region on a substrate that contains many copies of a molecule or molecules (*e.g.*, nucleic acid probes).

Each unit is composed of multiple cells that may be placed in rows (or "lanes") and columns. In one embodiment, a set of five related cells includes the following: a wild-type cell 220, "mutation" cells 222, and a "blank" cell 224. Cell 220 contains a wild-type probe that is the complement of a portion of the wild-type sequence. Cells 222 contain "mutation" probes for the wild-type sequence. For example, if the wild-type probe is 3'-ACGT, the probes 3'-ACAT, 3'-ACCT, 3'-ACGT, and 3'-ACTT may be the "mutation" probes. Cell 224 is the "blank" cell because it contains no probes (also called the "blank" probe). As the blank cell contains no probes, labeled targets should not bind to the chip in this area. Thus, the blank cell provides an area that can be used to measure the background intensity.

Again referring to Fig. 4, at step 206 the masks for the synthesis are designed. At step 208 the software utilizes the mask design and layout information to make the DNA or other polymer chips. This software 208 will control, among other things, relative translation of a substrate and the mask, the flow of desired reagents through a flow cell, the synthesis temperature of the flow cell, and other parameters. At step 210, another piece of software is used in scanning a chip thus synthesized and exposed to a labeled target. The software controls the scanning of the chip, and stores the data thus obtained in a file that may

later be utilized to extract sequence information.

At step 212 a computer system utilizes the layout information and the fluorescence information to evaluate the hybridized nucleic acid probes on the chip. Among the important pieces of information obtained from DNA chips are the identification of mutant targets and determination of genetic sequence of a particular target.

Fig. 6 illustrates the binding of a particular target DNA to an array of DNA probes 114. As shown in this simple example, the following probes are formed in the array (only one probe is shown for the wild-type probe):

3'-AGAACGT
AGACCGT
AGAGCGT
AGATCGT
.
.
.

As shown, the set of probes differ by only one base, a single base mismatch at an interrogation position, so the probes are designed to determine the identity of the base at that location in the nucleic acid sequence. Accordingly, when used herein a unit will refer to multiple sets of related probes, where each set includes probes that differ by a single base mismatch at an interrogation position.

When a fluorescein-labeled (or other marked) target with the sequence 5'-TCTTGCA is exposed to the array, it is complementary only to the probe 3'-AGAACGT, and fluorescein will be primarily found on the surface of the chip where 3'-AGAACGT is located. Thus, for each set of probes that differ by only one base, the image file will contain four fluorescence intensities, one for each probe. Each fluorescence intensity can therefore be associated with the nucleotide or base of each probe that is different from the other probes. Additionally, the image file will contain a "blank" cell which can be used as the fluorescence intensity of the background. By analyzing the five fluorescence intensities associated with a

specific base location, it becomes possible to extract sequence information from such arrays using the methods of the invention disclosed herein.

Fig. 7 illustrates probes arranged in lanes on a chip. A reference sequence (or chip wild-type sequence) is shown with five interrogation positions marked with number subscripts. An interrogation position is oftentimes a base position in the reference sequence where the target sequence may contain a mutation or otherwise differ from the reference sequence. The chip may contain five probe cells that correspond to each interrogation position. Each probe cell contains a set of probes that have a common base at the interrogation position. For example, at the first interrogation position, I_1, the reference sequence has a base T. The wild-type probe for this interrogation position is 3'-TGAC where the base A in the probe is complementary to the base at the interrogation position in the reference sequence.

Similarly, there are four "mutant" probe cells for the first interrogation position, I_1. The four mutant probes are 3'-TGAC, 3'-TGCC, 3'-TGGC, and 3'-TGTC. Each of the four mutant probes vary by a single base at the interrogation position. As shown, the wild-type and mutant probes are arranged in lanes on the chip. One of the mutant probes (in this case 3'-TGAC) is identical to the wild-type probe and therefore does not evidence a mutation. However, the redundancy gives a visual indication of mutations as will be seen in Fig. 8.

Still referring to Fig. 7, the chip contains wild-type and mutant probes for each of the other interrogation positions I_2-I_5. In each case, the wild-type probe is equivalent to one of the mutant probes.

Fig. 8 illustrates a hybridization pattern of a target on a chip with a reference sequence as in Fig. 7. The reference sequence is shown along the top of the chip for comparison. The chip includes a WT-lane (wild-type), an A-lane, a C-lane, a G-lane, and a T-lane (or U). Each lane is a row of cells containing probes. The cells in the WT-lane contain probes that are complementary to the reference sequence. The cells in the A-, C-, G-, and T-lanes contain probes that are complementary to the reference sequence except that the named base is at the interrogation position.

In one embodiment, the hybridization of probes in a cell is determined

by the fluorescent intensity (*e.g.*, photon counts) of the cell resulting from the binding of marked target sequences. The fluorescent intensity may vary greatly among cells. For simplicity, Fig. 8 shows a high degree of hybridization by a cell containing a darkened area. The WT-lane allows a simple visual indication that there is a mutation at interrogation position I_6 because the wild-type cell is not dark at that position. The cell in the C-lane is darkened which indicates that the mutation is from T->G (mutant probe cells are complementary so the C-cell indicates a G mutation). In a preferred embodiment, the WT-Lane is not utilized so four cells (not including any "blank" cell) are utilized to call a base at an interrogation position.

In practice, the fluorescent intensities of cells near an interrogation position having a mutation are relatively dark creating "dark regions" around a mutation. The lower fluorescent intensities result because the cells at interrogation positions near a mutation do not contain probes that are perfectly complementary to the target sequence; thus, the hybridization of these probes with the target sequence is lower. For example, the relative intensity of the cells at interrogation positions I_5 and I_9 may be relatively low because none of the probes therein are complementary to the target sequence. Although the lower fluorescent intensities reduce the resolution of the data, the methods of the present invention provide highly accurate base calling within the dark regions around a mutation and are able to identify other mutations within these regions.

Fig. 9 illustrates standard and alternate tilings on a chip. As shown, the chip includes twelve units (units$_{1-12}$). Units$_{1-4}$ are tiled (*i.e.*, designed and synthesized on the chip) to include probes complementary to the same reference sequence. For identification purposes, this group of units will be called the standard group. In general, base calls for the target sequence will be performed utilizing the standard group unless the invention determines that another group or groups should be utilized.

Units$_{5-8}$ are tiled to include probes complementary to the same reference sequence, but a reference sequence that differs from the reference sequence for the standard group. This group of units will be called an alternate group. Units$_{9-12}$ comprises another alternate group that are based on a reference

sequence that is different from the reference sequences of the standard and first alternate groups. Although the reference sequences are different, they are often quite similar. For example, the reference sequences may be slightly different mutations of HIV. Embodiments of the present invention evaluate and utilize information from tilings based on reference sequences that would typically not be used in base calling the target sequence.

The units within a group may include identical probes, probes of different structure, probes from the same or different chips, and the like. For example, one unit may include 5-mer probes with the interrogation position at the third position in probes. Another unit may include 10-mer probes with an interrogation position at the sixth position. Additionally, these units may have been tiled on the same or different chips.

The expanded section at the bottom left portion of Fig. 9 illustrates that each block of a unit typically includes four cells, denoted A, C, G, and T. The base designations specify which base is at the interrogation position of each probe within the cell. Typically, there are hundreds or thousands of identical nucleic probes within each cell.

Although in preferred embodiments the cells may be arranged adjacent to each other in sequential order along the reference sequence, there is no requirement that the cells be in any particular location as long as the location on the chip is determinable. Additionally, although it may be beneficial to synthesize the different groups on a single chip for consistency of experiments, the methods of the present invention may be advantageously utilized with data from different tilings on different chips.

Analyzing Target Sequences

Fig. 10 shows a screen display of hybridization intensities from a chip. During analysis, the system receives an image file including the scanned image of the hybridized chip. In a preferred embodiment, the image file shows fluorescent intensities and locations that labeled target nucleic acid sequences or fragments bound to the chip.

A screen display 260 utilizes the common windowing graphical user

interface. The user may select to display the image file for inspection. After the user selects the image file to be displayed, a window 262 is displayed that includes the image file. The image file shown includes multiple rows of A-, C-, G-, and T-lanes.

As the user moves the cursor over the displayed image file, a status bar 264 indicates the X and Y position of the cursor and the fluorescent intensity at that position. Additionally, the user is able to utilize the pointing device to select a rectangular area of the image file in order to manipulate the sub-image. For example, the user may magnify the subimage so that the individual cells may be seen more clearly. Additionally, the user may adjust the contrast of the intensities to bring to light some differences in hybridization intensity that is not apparent at the current contrast setting.

Fig. 11 is a flowchart of a process of computing a base call from hybridization intensities of related probes. When used herein, "related probes" are probes that differ by a nucleotide base at an interrogation position. Although typically the probes are identical except at the interrogation position, the probes may differ at other base positions as well. Accordingly, the related probes differ by at least one base.

At step 302 the hybridization intensities of the four related probes are adjusted by subtracting the background or "blank" cell intensity. Preferably, if a hybridization intensity is then less than or equal to zero, the hybridization intensity is set equal to a small positive number to prevent division by zero or negative numbers in future calculations.

At step 304, the hybridization intensities are sorted by intensity. The highest intensity is then compared to a predetermined background difference cutoff at step 306. The background difference cutoff is a number that specifies the hybridization intensity the highest intensity probe must be over the background intensity in order to correctly call the unknown base. Thus, the background adjusted base intensity must be greater than the background difference cutoff or the unknown base is deemed to be not accurately callable.

If the highest hybridization intensity of the related probes is not greater than the background difference cutoff, the unknown base is assigned the

code 'N' (insufficient intensity) as shown at step 308. Otherwise, the ratio of the highest hybridization intensity and second highest hybridization intensity is calculated as shown at step 310.

At step 312, the ratio calculated at step 310 is compared to a predetermined ratio cutoff. The ratio cutoff is a number that specifies the ratio required to identify the unknown base. In preferred embodiments, the ratio cutoff if 1.2. If the ratio is greater than the ratio cutoff, the unknown base is called according to the probe with the highest hybridization intensity. Typically, the base is called as the complement of the base at the interrogation position in the highest intensity probe as shown at step 314. Otherwise, the ratio of the second highest hybridization intensity and third highest hybridization intensity is calculated as shown at step 316.

At step 318, the ratio calculated at step 316 is compared to the ratio cutoff. If the ratio is greater than the ratio cutoff, the unknown base is called as being an ambiguity code specifying the complements of interrogation position bases of the highest hybridization intensity probe and the second highest hybridization probe as shown at step 320. Otherwise, the ratio of the third highest hybridization intensity and fourth highest hybridization intensity is calculated as shown at step 322.

At step 324, the ratio calculated at step 322 is compared to the ratio cutoff. If the ratio is greater than the ratio cutoff, the unknown base is called as being an ambiguity code specifying the complements of interrogation position bases of the highest, second highest and third highest hybridization intensity probes as shown at step 326. Otherwise, the unknown base is assigned the code 'X' (insufficient discrimination) as shown at step 328.

Fig. 12 is a flowchart of another process of computing a base call from hybridization intensities of related probes. The flowchart shown operates on hybridization intensities demonstrated by related probes; thus, a base call is made for the base in the target corresponding to the interrogation position in probes that differ by a single base mismatch at the interrogation position. At step 402, the system determines if there is one probe with the highest hybridization to the target sequence. If there is not, the base is called as an 'N' meaning ambiguous. For

example, if two probes have the same highest intensity (*i.e.*, there is a tie), the base would be called as 'N'.

If there is a single probe that has the highest hybridization to the target, the base is called according to that probe at step 406. Since the probes are complementary to the target sequence, the base may be called as the complementary base (C/G, A/T) to the base at the interrogation position of the probe.

At step 408, the system determines if the base call is a mutant, meaning it is different than the base in the reference sequence. If the base call is not a mutant base call, the base call has been made. Otherwise, the system determines checks to make sure certain "mutant" conditions are met at step 410 or the base is called as 'N' at step 412.

Before describing the mutant conditions for one embodiment, it may be beneficial to give labels to the hybridization intensities of the related probes. For illustration purposes "HighInt" will refer to the highest hybridization intensity, "SecondInt" will refer to the second highest hybridization intensity, "ThirdInt" will refer to the third highest hybridization intensity, and "LowInt" will refer to the lowest highest hybridization intensity.

In one embodiment, the mutant conditions include three tests that must all be met to call the base a mutant. A first test is whether the different between HighInt and SecondInt is greater than a difference cutoff. Thus, the system determines if HighInt - SecondInt is greater than a predefined value. This value should be chosen to allow mutant base calls only when the highest hybridization intensity is greater than the next highest hybridization intensity by a desired amount.

A second test is whether a first ratio is less than a first ratio cutoff. The first ratio is the following:

$$\frac{\text{SecondInt} - \text{sqrt}(\text{ThirdInt} * \text{LowInt})}{\text{HighInt} - \text{sqrt}(\text{ThirdInt} * \text{LowInt})}$$

The system determines if this first ratio is less than a predefined value. This value should be chosen to allow mutant base calls only when the highest hybridization intensity is a desired ratio greater than the next highest hybridization intensity even after the lowest two hybridization intensities are subtracted out.

A third test is whether a neighbor ratio is greater than a neighbor ratio cutoff. The neighbor ratio is the following:

$$\frac{HighInt_n}{HighInt_n - sqrt(HighInt_{n+1} * HighInt_{n-1})}$$

where the subscript n designates values for the base position that is being called and n+1 and n-1 represent values for adjacent base positions. Thus, the system determines if the neighbor ratio is greater than a predefined value. This value should be chosen to allow mutant base calls only when the highest hybridization intensity is a desired ratio greater than the highest hybridization intensity with the adjacent highest hybridization intensities subtracted out.

Accordingly, in a preferred embodiment, only if all of the mutant conditions are met will the base be called a mutant base. This embodiment recognizes that mutations are fairly rare so a mutant base should only be called when there is a high likelihood that there has been a mutation. If the mutant conditions are not met, the base may be called as ambiguous or as the same as the reference sequence (which statistically may be the correct base call).

Although a preferred embodiment utilizes three mutant conditions, other embodiments may use a single mutant condition (*e.g.*, one of the conditions described above). Other embodiments may utilize other base calling methods including the ones described in the U.S. Patent Applications previously incorporated by reference.

Fig. 13 is a flowchart of a process of calling bases in a group of units. As indicated earlier, a unit includes multiple sets of related cells, where the related cells include probes that differ by a single base at an interrogation position. In a typical embodiment, the system initially receives input on the hybridization intensities (*e.g.*, from the image data file produced by a scanner that scans the hybridized chip) and the structure of the probes that correspond to the hybridization intensities. In preferred embodiments, the background intensity (*e.g.*, intensity measured from "blank" cells or other areas of the chip without probes) are subtracted from the measured hybridization intensities. The background subtracted hybridization intensities may also be limited to have a minimum hybridization intensity of 1 (*e.g.*, one photon count).

The hybridization intensity describes the extent of hybridization that was measured between a probe (or multiple copies of a probe) and the target sequence. As an example, the hybridization intensity may refer to the mean of the photon counts recorded from a cell, the photon counts resulting from fluorescein labeled target sequences that bound to probes in the cell.

At step 452, the system gets a base position in the target sequence to be called. The system then computes a base call for each unit of the group at step 454. Therefore, the hybridization intensities for the related cells of each unit at the base position are analyzed. With this analysis (embodiments of which were described in more detail in reference to Figs. 11 and 12), the system computes a base call for each unit. Thus, if there are five units in the group, five base calls may be produced.

The system analyzes the base calls of the units of the group at step 456 in order to compute a base call for the group. In one embodiment, the system calls the base according to the base which is called most often by the units. For example, if there are five units and the following base calls were made for each unit:

'T' - three units

'G' - one unit

'N' - one unit

The base will be called a T since three out of five units agree. Ties may be broken by analyzing other factors like the highest average hybridization intensity of the unit or units that call each base in the tie. In a preferred embodiment, the invention utilizes the process described in Fig. 15.

At step 458, it is determined whether there is next base position to analyze. The present invention may be utilized to call all the bases of a target nucleic acid sequence so the process may, in effect, "walk" through the base positions. Additionally, the invention may be utilized to call only certain base positions (*e.g.*, mutation positions) so the process may skip certain base positions altogether.

Fig. 14 is a flowchart of a process of calling bases for multiple groups of units. As shown in Fig. 9, there may be multiple groups on one or more chips that are to be analyzed. The multiple groups may be tiled according to different

reference sequences; however, this does not mean that all of their hybridization information may not be utilized. Typically, it is assumed that the reference sequence for the standard group is expected to be the most identical to the target sequence. However, if one of the alternate groups is determined to be more identical (*i.e.*, better for making a base call), then that group will be used to make the base call.

At step 502, the system computes base calls in the units of the standard and alternate groups. The base calling may be done as was described in reference to Fig. 13.

The system then computes a base call for each group of units at step 504. This may be accomplished by determining the base that is called most often by the units. Alternatively, the base call for the group may be determined utilizing the process which will be described in more detail in reference to Fig. 15.

After the system has determined a base call for each group of units (both the standard and alternate tilings), the system identifies a base position at step 506. The system then determines the best group of units for this base position to be utilized to make the base call. In general, selecting the best group may involve determining which reference sequence of the groups has the fewest mismatches with the target sequence near or in a window around the interrogation position. The group of units that has the fewest mismatches near the interrogation position may have the highest likelihood of producing the most accurate base call. An embodiment of selecting the best group will be described in more detail in reference to Fig. 16.

At step 510, the system calls the base at the identified base position according to the best group of units (*i.e.*, utilizing the base call for the group that was computed at step 504). Once the base call has been made, the system determines if there is a next base position to perform a base call. If there is another base position to be called, the system proceeds to call that base position at step 506.

Fig. 15 is a flowchart of a process of calling a base for a group of units. At step 602, the system determines if a majority of units call the same base at the specified base position. The majority is determined upon reference to only those units that call a base (*e.g.*, do not call as ambiguous or 'N'). For example, assume

that there are seven units and the following base calls have been made for the units:

'G' - three units

'T' - one unit

'N' - four units

Since three out of four of the nonambiguous base calls are 'G', the system will initially call the base as a 'G' for the group of units. The base will be called as the majority base unless an exception rule applies at step 604.

The exception rules specify conditions which dictate what base call should be made for the group of units. These rules may include conditions that change a majority base call and may include conditions to deal with situations when there is not a base call that a majority of units call. In a preferred embodiment, the exception rules include tie breaking rules which analyze the hybridization intensity of neighboring probes (*e.g.*, one unit calls one base and another unit calls a different base). Additionally, the exception rules specify that if three units call different bases with one of the calls being for the reference base, the system should call the base as the reference for the group of units. Other exception rules are described in the Appendix.

At step 606, the system determines if an exception rule applies. If an exception rule does apply, the rule is applied at step 608.

Fig. 16 is a flowchart of a process of selecting a best group of units for performing a base call. Selecting the best group involves determining which reference sequence of the groups has the fewest mismatches with the target sequence near the interrogation position. The group of units that has the fewest mismatches near the interrogation position may have the highest likelihood of producing the most accurate base call. The window around the interrogation position which is analyzed may be a set value or set according to the probe structure. For example, if the maximum distance that the probes for all the groups extend from the interrogation position is eight base positions to one side of the interrogation position and ten base positions to the other side of the interrogation position, the window may be set as including this range of base positions.

At step 702, the system calculates mismatch scores for the standard and alternate groups of units. The mismatch score is an indication of how many

mismatches a reference sequence appears to have with the target sequence. In order to determine a mismatch score, the system may only analyze base positions where at least two of the reference sequences differ. Thus, if all the reference sequences are identical at a base position, this base position may be skipped.

At each base position where at least two reference sequences differ, the system determines if the base call for a group (the base call indicating the likely base in the target sequence) at each of these positions differs from the corresponding base of the reference sequence. If the base call and the base for the reference sequence differ, the mismatch score is incremented by one. Initially, the mismatch scores for each group is set to zero.

Conceptually, it should be understood that the mismatch score is an indication of the number of base positions in a portion of the reference sequence that differ from the target sequence (optionally excluding those positions where all the reference sequences are the same). To better illustrate this concept, the following simple example is presented. Assume there is a standard group and two alternate groups as follows:

	Standard Group	Mismatch Score
reference	ACGGATGAGATACGA	1
base calls	ACTGATGAGATACGA	

	Alternate Group 1	Mismatch Score
reference	ACTGATGAGATACGA	0
base calls	ACTGATGAGATACGA	

	Alternate Group 2	Mismatch Score
reference	ACGGATGAGATACGT	2
base calls	ACTGATGAGATACGA	

The underlined bases correspond to the base position which is being analyzed. The bolded base positions indicate base positions where at least two of the reference sequences differ. At these bolded base positions, the standard group has one base position where the reference sequence differs from the target sequence (as indicated by the base calls) so the mismatch score is 1. Similarly, the first alternate group has

a mismatch score of 0 and the second alternate group has a mismatch score of 2.

As alternate group 1 has the lowest mismatch score, that group would be utilized to call the base at the base position being analyzed. In this simple example, the base call is not different for any of the groups as this example is intended to illustrate how the best group may be selected. However, what is important is that the invention recognizes that the more mismatches that occur near a base position, the less accurate the base call will become. This result is brought upon by the fact that a mismatch between the reference sequence and the target sequence creates any area where the probes interrogating neighboring base positions include a single base mismatch. Single base mismatches lower the hybridization intensity and may produce inaccurate results.

At step 704, the system determines if a mismatch score of the standard groups is less than or equal to the mismatch scores of alternate groups. If the standard group has the lowest mismatch score (or ties), then the base call performed according to the standard group.

The system determines if a single alternate group has the lowest mismatch score at step 708. If so, that alternate group is utilized to make the base call at step 710. Otherwise, there are more than one alternate groups that have the same mismatch scores. If this is the case, the alternate group may be chosen which includes units that most consistently called the base at step 712. For example, if two alternate groups have the same lowest mismatch score but one group's units all called the same base and the other group's units were split, the alternate group that called the same base would be utilized. Other methods of determining the best group in the event of a mismatch score tie may also be utilized.

Fig. 17A shows a screen displays allowing analysis of nucleotides from experiments from one or more chips. A screen display 802 includes multiple screen areas that display different information to the user. A screen area 804 includes the name of a reference sequence which in this example is PRT 440A which are antisense regions (Protease Reverse Transcriptase) of the HIV virus. The reference sequence is typically used as a baseline with which to compare sample sequences. Although the reference sequence on the screen may be the chip wild-type sequence for which the chips were tiled, there is no requirement that this is the

case.

A screen area 806 includes the nucleotide sequence for the reference sequence for the probe array. The base position of each nucleotide is shown above screen area 806. Screen area 806 also sohows the reference sequence for each unit if "expanded" in the user interface.

A screen area 808 shows the user the chip and composite files that are currently being analyzed. A chip file (*e.g.*, ends in ".CHP") includes data obtained from a single chip. A composite file (*e.g.*, ends in ".CMP") includes data obtained from more than one chip. When a user opens a chip or composite file for analysis, the pathname of the file is displayed in screen area 808.

Information from the chip and composite files may be displayed in screen areas 810 and 812. Screen area 810 includes the names of sample sequences currently being analyzed from the chip or composite files. The name of the sample sequence is typically chosen to enable the user to readily determine the what the sample sequence represents. Screen area 812 includes the nucleotide sequence for the sample sequences. The base position of each nucleotide in screen area is the same as indicated above screen area 806. Accordingly, the system automatically aligns the reference and sample sequences for easier analysis.

Fig. 17A has been described in order to familiarize the reader with the layout of the screen display. However, as illustrated by Fig. 17B, the invention allows the user to hide (not display) and summarize information from chip and composite files. For example, if a user "clicks on" or activates the screen icon plus sign in front of the composite filename in screen area 808, the system displays more information about the composite file. As shown, the method that was utilized to combine the information from the chip files may be shown along with the individual chip files.

Additionally, if a user activates the screen icon plus sign in front of the chip filename in screen area 808, the system displays more information about the chip file including the process or procedure that was utilized to calls bases. In Fig. 17B, the base calling procedure was the "Ratio Base Algorithm" which was described in reference to Fig. 10. Additionally, the user is able to modify parameters for the base calling procedure which will be immediately reflected in the

base calls shown on the display screen. For example, the ratio cutoff ("Ratio") is displayed as 1.2. If a user increases the ratio cutoff to 1.4, the system would then recalculate the base calls for the chip and the new base calls would be reflected in screen area 812. The parameters may be any values the are input into the base calling procedure including constants, thresholds, ranges, and the like.

Fig. 17B also illustrates that the system is able to combine data from multiple experiments (including various filings) for easier reading of the user. The sample sequence 440-2A was shown in Fig. 17A and has been expanded in Fig. 17B to show that the base calls are derived from multiple experiments, where the data from multiple experiments may be from a single chip or multiple chips. In other words, the nucleotide sequence shown for sample sequence 440-2A in Figs. 17A and 17B do not represent a single experiment but actually a combination or consensus from multiple experiments. The user is able to review the data from each of the experiments as shown in Fig. 17B which includes displaying the hybridization intensities for each related base. The system allows the user to highlight a base position for analysis as shown for base position 100.

Referring again to Fig. 17A, a screen icon plus sign is displayed in front of the name of the sample sequence "440-2A." By activating the screen icon, the system displays each of the individual calls that make up the composite base call. As shown in Fig. 17B, the composite base call is derived from multiple base calls. The multiple base calls are aligned with the composite base call according to base position. The invention provides great flexibility to the user for displaying, hiding, and summarizing data for analyzing sequences.

Monitoring Gene Expression

Fig. 18 shows a high level flowchart of a process of monitoring the expression of a gene by comparing hybridization intensities of pairs of perfect match and mismatch probes. The term "perfect match probe" refers to a probe that has a sequence that is perfectly complementary to a particular target sequence. The test probe is typically perfectly complementary to a portion (subsequence) of the target sequence. The term "mismatch control" or "mismatch probe" refer to probes whose sequence is deliberately selected not to be perfectly complementary to

a particular target sequence. For each mismatch (MM) control in a high-density array there typically exists a corresponding perfect match (PM) probe that is perfectly complementary to the same particular target sequence.

The process compares hybridization intensities of pairs of perfect match and mismatch probes that are preferably covalently attached to the surface of a substrate or chip. Most preferably, the nucleic acid probes have a density greater than about 60 different nucleic acid probes per 1 cm^2 of the substrate. Although the flowcharts show a sequence of steps for clarity, this is not an indication that the steps must be performed in this specific order. One of ordinary skill in the art would readily recognize that many of the steps may be reordered, combined, and deleted without departing from the invention.

Initially, nucleic acid probes are selected that are complementary to the target sequence (or gene). These probes are the perfect match probes. Another set of probes is specified that are intended to be not perfectly complementary to the target sequence. These probes are the mismatch probes and each mismatch probe includes at least one nucleotide mismatch from a perfect match probe. Accordingly, a mismatch probe and the perfect match probe from which it was derived make up a pair of probes. As mentioned earlier, the nucleotide mismatch is preferably near the center of the mismatch probe.

The probe lengths of the perfect match probes are typically chosen to exhibit high hybridization affinity with the target sequence. For example, the nucleic acid probes may be all 20-mers. However, probes of varying lengths may also be synthesized on the substrate for any number of reasons including resolving ambiguities.

The target sequence is typically fragmented, labeled and exposed to a substrate including the nucleic acid probes as described earlier. The hybridization intensities of the nucleic acid probes is then measured and input into a computer system. The computer system may be the same system that directs the substrate hybridization or it may be a different system altogether. Of course, any computer system for use with the invention should have available other details of the experiment including possibly the gene name, gene sequence, probe sequences, probe locations on the substrate, and the like.

Referring to Fig. 18, after hybridization, the computer system receives input of hybridization intensities of the multiple pairs of perfect match and mismatch probes at step 902. The hybridization intensities indicate hybridization affinity between the nucleic acid probes and the target nucleic acid (which corresponds to a gene). Each pair includes a perfect match probe that is perfectly complementary to a portion of the target nucleic acid and a mismatch probe that differs from the perfect match probe by at least one nucleotide.

At step 904, the computer system compares the hybridization intensities of the perfect match and mismatch probes of each pair. If the gene is expressed, the hybridization intensity (or affinity) of a perfect match probe of a pair should be recognizably higher than the corresponding mismatch probe. Generally, if the hybridizations intensities of a pair of probes are substantially the same, it may indicate the gene is not expressed. However, the determination is not based on a single pair of probes, the determination of whether a gene is expressed is based on an analysis of many pairs of probes. An exemplary process of comparing the hybridization intensities of the pairs of probes will be described in more detail in reference to Fig. 19.

After the system compares the hybridization intensity of the perfect match and mismatch probes, the system indicates expression of the gene at step 906. As an example, the system may indicate an expression call to a user that the gene is either present (expressed), marginal or absent (unexpressed).

Fig. 19 shows a flowchart of a process of determining if a gene is expressed utilizing a decision matrix. At step 952, the computer system receives raw scan data of N pairs of perfect match and mismatch probes. In a preferred embodiment, the hybridization intensities are photon counts from a fluorescein labeled target that has hybridized to the probes on the substrate. For simplicity, the hybridization intensity of a perfect match probe will be designed "I_{pm}" and the hybridization intensity of a mismatch probe will be designed "I_{mis}."

Hybridization intensities for a pair of probes is retrieved at step 954. The background signal intensity is subtracted from each of the hybridization intensities of the pair at step 956. Background subtraction may also be performed on all the raw scan data at the same time.

At step 958, the hybridization intensities of the pair of probes are compared to a difference threshold (D) and a ratio threshold (R). It is determined if the difference between the hybridization intensities of the pair ($I_{pm} - I_{mm}$) is greater than or equal to the difference threshold AND the quotient of the hybridization intensities of the pair (I_{pm} / I_{mm}) is greater than or equal to the ratio threshold. The difference thresholds are typically user defined values that have been determined to produce accurate expression monitoring of a gene or genes. In one embodiment, the difference threshold is 20 and the ratio threshold is 1.2.

If $I_{pm} - I_{mm} >= D$ and $I_{pm} / I_{mm} >= R$, the value NPOS is incremented at step 960. In general, NPOS is a value that indicates the number of pairs of probes which have hybridization intensities indicating that the gene is likely expressed. NPOS is utilized in a determination of the expression of the gene.

At step 962, it is determined if $I_{mm} - I_{pm} >= D$ and $I_{mm} / I_{pm} >= R$. If this expression is true, the value NNEG is incremented at step 964. In general, NNEG is a value that indicates the number of pairs of probes which have hybridization intensities indicating that the gene is likely not expressed. NNEG, like NPOS, is utilized in a determination of the expression of the gene.

For each pair that exhibits hybridization intensities either indicating the gene is expressed or not expressed, a log ratio value (LR) and intensity difference value (IDIF) are calculated at step 966. LR is calculated by the log of the quotient of the hybridization intensities of the pair (I_{pm} / I_{mm}). The IDIF is calculated by the difference between the hybridization intensities of the pair ($I_{pm} - I_{mm}$). If there is a next pair of hybridization intensities at step 968, they are retrieved at step 954.

At step 972, a decision matrix is utilized to indicate if the gene is expressed. The decision matrix utilizes the values N, NPOS, NNEG, and LR (multiple LRs). The following four assignments are performed:

P1 = NPOS / NNEG

P2 = NPOS / N

P3 = (10 * SUM(LR)) / (NPOS + NNEG)

These P values are then utilized to determine if the gene is expressed.

For purposes of illustration, the P values are broken down into ranges.

If P1 is greater than or equal to 2.1, then A is true. If P1 is less than 2.1 and greater than or equal to 1.8, then B is true. Otherwise, C is true. Thus, P1 is broken down into three ranges A, B and C. This is done to aid the readers understanding of the invention.

Thus, all of the P values are broken down into ranges according to the following:

A = (P1 >= 2.1)

B = (2.1 > P1 >= 1.8)

C = (P1 < 1.8)

X = (P2 >= 0.35)

Y = (0.35 > P2 >= 0.20)

Z = (P2 < 0.20)

Q = (P3 >= 1.5)

R = (1.5 > P3 >= 1.1)

S = (P3 < 1.1)

Once the P values are broken down into ranges according to the above boolean values, the gene expression is determined.

The gene expression is indicated as present (expressed), marginal or absent (not expressed). The gene is indicated as expressed if the following expression is true: A and (X or Y) and (Q or R). In other words, the gene is indicated as expressed if P1 >= 2.1, P2 >= 0.20 and P3 >= 1.1. Additionally, the gene is indicated as expressed if the following expression is true: B and X and Q.

With the forgoing explanation, the following is a summary of the gene expression indications:

Present	A and (X or Y) and (Q or R)
	B and X and I
Marginal	A and X and S
	B and X and R
	B and Y and (Q or R)

Absent All others cases (*e.g.*, any C combination)

In the output to the user, present may be indicated as "P," marginal as "M" and absent as "A" at step 974.

Once all the pairs of probes have been processed and the expression of the gene indicated, an average of ten times the LRs is computed at step 975. Additionally, an average of the IDIF values for the probes that incremented NPOS and NNEG is calculated, which may be utilized as an expression level. These values may be utilized for quantitative comparisons of this experiments with other experiments.

Quantitative measurements may be performed at step 976. For example, the current experiment may be compared to a previous experiment (*e.g.*, utilizing values calculated at step 970). Additionally, the experiment may be compared to hybridization intensities of RNA (such as from bacteria) present in the biological sample in a known quantity. In this manner, one may verify the correctness of the gene expression indication or call, modify threshold values, or perform any number of modifications of the preceding.

For simplicity, Fig. 19 was described in reference to a single gene. However, the process may be utilized on multiple genes in a biological sample. Therefore, any discussion of the analysis of a single gene is not an indication that the process may not be extended to processing multiple genes.

Fig. 20 shows a screen display layout of gene expression monitoring software. A screen display 1000 is divided into two sections: a graphics display area 1002 and a data display area 1004. The graphics display area is for displaying graphs which will aid the user in interpreting the data. The data display area is for displaying the underlying data so the user may evaluate the underlying data for gene expression.

As will be shown in subsequent screen displays, the data display area is preferably organized in a table having rows and columns. Each column has a heading indicating the data that resides in the column. Each row represents data from a single experiment or combination of experiments for a gene. The term "experiment" is used herein to describe a process that created data. For example, a

single image file of a hybridized chip may produce many "experiments" for a number of genes. Additionally, experiments may refer to data obtained from different chips.

Fig. 21A shows a screen display illustrating the analysis of a selected gene. A screen display 1030 includes a graphics display area that illustrates with bar graphs the hybridization intensities of perfect match probes and mismatch probes at each base position in a selected gene. The gene selected is shown highlighted in a data display area 1034.

The data display area includes a number of column headings. The Experiment Name refers to a user-defined name for the experiment. The Gene Name is the name of the gene. The numbers Positive and Negative refer to the values NPOS and NNEG as described in reference to Fig. 19. The Pairs column indicates the number of perfect match and mismatch probe pairs that were utilized in the analysis of the gene. The Pos Fraction column indicates the fraction of probe pairs that were scored as positive (*i.e.*, Positive/Pairs).

The Avg Ratio column indicates the average of I_{pm}/I_{mm} for all probes for a gene. The Log Avg column indicates the average of the $\log(I_{pm}/I_{mm})$. The PM Excess column indicates the number of perfect match probes that exhibit a hybridization intensity above a user defined threshold. The MM Excess indicates the number of mismatch probes that exhibit a hybridization intensity above a user defined threshold. Referring now to Fig. 21B, the Pos/Neg column indicates ratio of the Positive column to the Negative column ("Inf" is utilized if the Negative column includes a zero). The Avg Diff column indicates the average intensity difference for the gene. The average intensity difference was computed at step 975 of Fig. 19 (*i.e.*, average(IDIF)).

The Abs Call column indicates the gene expression call for the experiment. The values in this column may be "P" for present, "M" for marginal and "A" for absent. The gene expression call for a preferred embodiment is described in more detail in reference to step 974 of Fig. 19.

As the user selects an experiment, the graphics display area displays graphs to aid the user in interpreting the data. A button bar 1034 enables the user to select which graph or graphs to display in the graphics display area.

Additionally, the user is able to sort the data in the display data are according to values in a selected column.

Fig. 22 shows another screen display illustrating the analysis of a selected gene. A screen display 1060 includes a graphics display area 1062 illustrating a graph of the ratio of the hybridization intensity of the perfect match probe to the mismatch probe at each base position. The x-axis is the base position and the y-axis is the ratio of hybridization intensities. The statistical ratio threshold is plotted on the graph, which in this example is 1.2. this graph may be utilized by the user to analyze how many probe pairs (I_{pm}/I_{mm}) are above or below the threshold. The graph also includes the gene and experiment names.

Fig. 23 shows a screen display illustrating the comparison of experiments for selected genes. A screen display 1160 includes a graphics display area 1062 and a data display are 1164. The graphics display area includes a graph of the ratio of the hybridization intensity of the perfect match probe to the mismatch probe at each base position for each of the experiments/genes selected in the data display area. In a preferred embodiment, the experiment name, gene name, and data plot are a different color for each gene to allow the user to more easily see the differences between or among selected genes.

Fig. 24 shows another screen display illustrating the comparison of experiments for selected genes. A screen display 1200 includes a graphics display area 1202 illustrating the expression levels of genes selected in a data display area 1204. The graph of the expression levels of the selected genes is a bar graph. In a preferred embodiment, the expression level is defined as the average intensity difference (*see* average(IDIF) in Fig. 19). The graph also includes the gene and experiment names.

Fig. 25 shows another screen display illustrating the comparison of experiments for selected genes with multiple graphs in the graphics display area. A screen display 1230 includes a graphics display area 1232 depicting multiple graphs for analyzing the genes selected in a data display area 1234. An expression level graph 1236, an average intensity difference graph 1238 and a hybridization intensity graph 1240 are shown for the selected genes.

Figs. 26A and 26B show the flow of a process of determining the

expression of a gene by comparing baseline scan data and experimental scan data. For example, the baseline scan data may be from a biological sample where it is known the gene is expressed. Thus, this scan data may be compared to a different biological sample to determine if the gene is expressed. Additionally, it may be determined how the expression of a gene or genes changes over time in a biological organism. Accordingly, the term "baseline" means that it will be used as a point of reference.

At step 1302, the computer system receives raw scan data of N pairs of perfect match and mismatch probes from the baseline. The hybridization intensity of a perfect match probe from the baseline will be designed "I_{pm}" and the hybridization intensity of a mismatch probe from the baseline will be designed "I_{mm}." The background signal intensity is subtracted from each of the hybridization intensities of the pairs of baseline scan data at step 1304.

At step 1306, the computer system receives raw scan data of N pairs of perfect match and mismatch probes from the experimental biological sample. The hybridization intensity of a perfect match probes from the experiment will be designed "J_{pm}" and the hybridization intensity of a mismatch probe from the experiment will be designed "J_{mm}." The background signal intensity is subtracted from each of the hybridization intensities of the pairs of experimental scan data at step 1308.

The hybridization intensities of an I and J pair may be normalized at step 1310. For example, the hybridization intensities of the I and J pairs may be divided by the hybridization intensity of control probes.

At step 1312, the hybridization intensities of the I and J pair of probes are compared to a difference threshold (DDIF) and a ratio threshold (RDIF). It is determined if the difference between the hybridization intensities of the one pair (J_{pm} - J_{mm}) and the other pair (I_{pm} - I_{mm}) are greater than or equal to the difference threshold AND the quotient of the hybridization intensities of one pair (J_{pm} - J_{mm}) and the other pair (I_{pm} - I_{mm}) are greater than or equal to the ratio threshold. The difference thresholds are typically user defined values that have been determined to produce accurate expression monitoring of a gene or genes.

If (J_{pm} - J_{mm}) - (I_{pm} - I_{mm}) >= DDIF and (J_{pm} - J_{mm}) / (I_{pm} - I_{mm}) >= RDIF,

the value NINC is incremented at step 1314. In general, NINC is a value that indicates the experimental pair of probes indicates that the gene expression is likely greater (or increased) than the baseline sample. NINC is utilized in a determination of whether the expression of the gene is greater (or increased), less (or decreased) or did not change in the experimental sample compared to the baseline sample.

At step 1316, it is determined if $(I_{pm} - I_{mm}) - (I_{pm} - I_{mm}) >= DDIF$ and $(I_{pm} - I_{mm}) / (I_{pm} / I_{mm}) >= RDIF$. If this expression is true, NDEC is incremented. In general, NDEC is a value that indicates the experimental pair of probes indicates that the gene expression is likely less (or decreased) than the baseline sample NDEC is utilized in a determination of whether the expression of the gene is greater (or increased), less (or decreased) or did not change in the experimental sample compared to the baseline sample.

For each of the pairs that exhibits hybridization intensities either indicating the gene is expressed more or less in the experimental sample, the values NPOS, NNEG and LR are calculated for each pair of probes. These values are calculated as discussed above in reference to Fig. 19. A suffix of either "B" or "E" has been added to each value in order to indicate if the value denotes the baseline sample or the experimental sample, respectively. If there are next pairs of hybridization intensities at step 1322, they are processed in a similar manner as shown.

Referring now to Fig. 26B, an absolute decision computation is performed for both the baseline and experimental samples at step 1324. The absolute decision computation is an indication of whether the gene is expressed, marginal or absent in each of the baseline and experimental samples. Accordingly, in a preferred embodiment, this step entails performing steps 972 and 974 from Fig. 19 for each of the samples. This being done, there is an indication of gene expression for each of the samples taken alone.

At step 1326, a decision matrix is utilized to determine the difference in gene expression between the two samples. This decision matrix utilizes the values, N, NPOSB, NPOSE, NNEGB, NNEGE, NINC, NDEC, LRB, and LRE as they were calculated above. The decision matrix performs different calculations

depending on whether NINC is greater than or equal to NDEC. The calculations are as follows.

If NINC >= NDEC, the following four P values are determined:

P1 = NINC / NDEC

P2 = NINC / N

P3 = ((NPOSE - NPOSB) - (NNEGE - NNEGB)) / N

P4 = 10 * SUM(LRE - LRB) / N

These P values are then utilized to determine the difference in gene expression between the two samples.

For purposes of illustration, the P values are broken down into ranges as was done previously. Thus, all of the P values are broken down into ranges according to the following:

A = (P1 >= 2.8)

B = (2.8 > P1 >= 2.0)

C = (P1 < 2.0)

X = (P2 >= 0.34)

Y = (0.34 > P2 >= 0.24)

Z = (P2 < 0.24)

M = (P3 >= 0.20)

N = (0.20 > P3 >= 0.12)

O = (P3 < 0.12)

Q = (P4 >= 0.9)

R = (0.9 > P4 >= 0.5)

S = (P4 < 0.5)

Once the P values are broken down into ranges according to the above boolean values, the difference in gene expression between the two samples is determined.

In this case where NINC >= NDEC, the gene expression change is indicated as increased, marginal increase or no change. The following is a

summary of the gene expression indications:

Increased	A and (X or Y) and (Q or R) and (M or N or O)
	A and (X or Y) and (Q or R or S) and (M or N)
	B and (X or Y) and (Q or R) and (M or N)
	A and X and (Q or R or S) and (M or N or O)
Marginal	A or Y or S or O
Increase	B and (X or Y) and (Q or R) and O
	B and (X or Y) and S and (M or N)
	C and (X or Y) and (Q or R) and (M or N)
No Change	All others cases (*e.g.*, any Z combination)

In the output to the user, increased may be indicated as "I," marginal increase as "MI" and no change as "NC."

If NINC < NDEC, the following four P values are determined:

P1 = NDEC / NINC

P2 = NDEC / N

P3 = ((NNEGE - NNEGB) - (NPOSE - NPOSB)) / N

P4 = 10 * SUM(LRE - LRB) / N

These P values are then utilized to determine the difference in gene expression between the two samples.

The P values are broken down into the same ranges as for the other case where NINC >= NDEC. Thus, P values in this case indicate the same ranges and will not be repeated for the sake of brevity. However, the ranges generally indicate different changes in the gene expression between the two samples as shown below.

In this case where NINC < NDEC, the gene expression change is indicated as decreased, marginal decrease or no change. The following is a summary of the gene expression indications:

Decreased	A and (X or Y) and (Q or R) and (M or N or O)
	A and (X or Y) and (Q or R or S) and (M or N)
	B and (X or Y) and (Q or R) and (M or N)
	A and X and (Q or R or S) and (M or N or O)
Marginal	A or Y or S or O
Decrease	B and (X or Y) and (Q or R) and O
	B and (X or Y) and S and (M or N)
	C and (X or Y) and (Q or R) and (M or N)
No Change	All others cases (*e.g.*, any Z combination)

In the output to the user, decreased may be indicated as "D," marginal decrease as "MD" and no change as "NC."

The above has shown that the relative difference between the gene expression between a baseline sample and an experimental sample may be determined. An additional test may be performed that would change an I, MI, D, or MD (*i.e.*, not NC) call to NC if the gene is indicated as expressed in both samples (*e.g.*, from step 1324) and the following expressions are all true:

Average(IDIFB) >= 200

Average(IDIFE) >= 200

1.4 >= Average(IDIFE) / Average(IDIFB) >= 0.7

Thus, when a gene is expressed in both samples, a call of increased or decreased (whether marginal or not) will be changed to a no change call if the average intensity difference for each sample is relatively large or substantially the same for both samples. The IDIFB and IDIFE are calculated as the sum of all the IDIFs for each sample divided by N.

At step 1328, values for quantitative difference evaluation are calculated. An average of $((I_{pm} - I_{mm}) - (I_{pm} - I_{mm}))$ for each of the pairs is calculated.

Additionally, a quotient of the average of I_{pm} - I_{mm} and the average of I_{pm} - I_{mm} is calculated. These values may be utilized to compare the results with other experiments in step 1330.

Fig. 27A shows a screen display illustrating the monitoring of the change of gene expression between experiments. A screen display 1400 includes a graphics display area 1402 and a data display area 1404. A user begins the comparison of experiments for a gene by selecting two experiments for a gene. For simplicity, we will call one baseline data and the other experimental data, meaning it will be compared to the baseline. For example, a user may select two experiments for the gene with the name "g182506." A comparison of two experiments is an experiment itself so the user is able to enter an experiment name which was entered as "foo" in the data display area of Fig. 27A. Fig. 27B shows another screen display illustrating monitoring of the change of gene expression between experiments.

The system then determines the change in gene expression between the selected experiments according to the process described in Figs. 28A and 28B. The data display area includes columns denoting the data produced by this comparison. The Experiment Name refers to a user-defined name for the comparison experiment. The Gene Name is the name of the gene. The numbers Inc and Dec refer to the values NINC and NDEC as described in reference to Fig. 26A. More specifically, Inc refers to the number of base positions in the gene for which the difference and ratio of the perfect match and mismatch hybridization intensities are significantly greater in the experimental data.

The Inc Ratio column indicates the number of base positions where the hybridization intensity increased divided by the total number of base positions in the gene which are analyzed. The Dec Ratio column indicates the number of base positions where the hybridization intensity decreased divided by the total number of base positions in the gene which are analyzed. The Pos Change column indicates the difference in the number of positive scoring probe pairs in the experimental data versus the baseline data. The Neg Change column indicates the difference in the number of negative scoring probe pairs (perfect match and mismatch) in the experimental data versus the baseline data.

The Inc/Dec column indicates the number probe pairs which had an increase in hybridization intensity in the experimental data versus the number of probe pairs which had a decrease in hybridization intensity in the experimental data. The Avg Diff column indicates the average intensity difference in the experimental data.

The Diff Call column (not shown) indicates the change in expression level between the experiments for the gene. The column shows a "I" for increased gene expression, "MI" for marginal increased gene expression, "D" for decreased gene expression, "MD" for marginal decreased gene expression, "NC" for no change, and "?" for unknown. In a preferred embodiment, the change in expression level is calculated as described in reference to step 1326 of Fig. 26B.

In addition to calculating the change in gene expression, the user may also select graphs to analyze the data. Graphics display area 1402 shows three different graphs depicting the data from the baseline and experimental data.

Fig. 28 shows a screen display illustrating a three-dimensional bar graph which illustrates the change of gene expression between experiments. A screen display 1440 displays a graphical display area 1442 including a three-dimensional bar graph of the expression level of selected genes in a data display area 1444. The user selects one or more genes in the data display area and then instructs the system to generate a three-dimensional bar graph of the expression level of these genes, where the expression level in a preferred embodiment is the average intensity difference (*i.e.*, average(IDIF). The three-dimensional bar graph allows the user to easily view the expression level of multiple genes. Additionally, similar genes slected from multiple experiments may be shown simultaneously and rotated to display differences in experssion levels.

Conclusion

The above description is illustrative and not restrictive. Many variations of the invention will become apparent to those of skill in the art upon review of this disclosure. Merely by way of example, while the invention is illustrated with particular reference to the evaluation of DNA (natural or unnatural), the methods can be used in the analysis from chips with other materials synthesized

thereon, such as RNA. The scope of the invention should, therefore, be determined not with reference to the above description, but instead should be determined with reference to the appended claims along with their full scope of equivalents.

4 Brief Description of Drawings

Fig. 1 illustrates an example of a computer system that may be used to execute software embodiments of the present invention;

Fig. 2 shows a system block diagram of a typical computer system;

Fig. 3 illustrates an overall system for forming and analyzing arrays of biological materials such as DNA or RNA;

Fig. 4 is an illustration of an embodiment of software for the overall system;

Fig. 5 illustrates the global layout of a chip formed in the overall system;

Fig. 6 illustrates conceptually the binding of nucleic acid probes on chips to a labeled target;

Fig. 7 illustrates nucleic acid probes arranged in lanes on a chip;

Fig. 8 illustrates a hybridization pattern of a target on a chip with a reference sequence as in Fig. 7;

Fig. 9 illustrates standard and alternate tilings;

Fig. 10 shows a screen display of hybridization intensities from a chip;

Fig. 11 is a flowchart of a process of computing a base call from hybridization intensities of related probes;

Fig. 12 is a flowchart of another process of computing a base call from hybridization intensities of related probes;

Fig. 13 is a flowchart of a process of calling bases in a group of units;

Fig. 14 is a flowchart of a process of calling bases for multiple groups of units;

Fig. 15 is a flowchart of a process of calling a base for a group of units;

Fig. 16 is a flowchart of a process of selecting a best group of units for performing a base call;

Figs. 17A and 17B show screen displays allowing analysis of

nucleotides from experiments from one or more chips;

Fig. 18 shows a high level flowchart of a process of monitoring the expression of a gene by comparing hybridization intensities of pairs of perfect match and mismatch probes;

Fig. 19 shows a flowchart of a process of determining if a gene is expressed utilizing a decision matrix;

Fig. 20 shows a screen display layout of gene expression monitoring software;

Figs. 21A and 21B show screen displays illustrating the analysis of a selected gene;

Fig. 22 shows another screen display illustrating the analysis of a selected gene;

Fig. 23 shows a screen display illustrating the comparison of experiments for selected genes;

Fig. 24 shows another screen display illustrating the comparison of experiments for selected genes;

Fig. 25 shows another screen display illustrating the comparison of experiments for selected genes with multiple graphs in the graphics display area;

Figs. 26A and 26B show a flowchart of a process of determining the expression of a gene by comparing baseline scan data and experimental scan data;

Figs. 27A and 27B show screen displays illustrating the monitoring of the change of gene expression between experiments; and

Fig. 28 shows a screen display illustrating a three-dimensional bar graph which illustrates the change of gene expression between experiments.